ENERGY LANDSCAPE FOR LARGE AVERAGE SUBMATRIX DETECTION PROBLEMS IN GAUSSIAN RANDOM MATRICES

SHANKAR BHAMIDI[1], PARTHA S. DEY[2], AND ANDREW B. NOBEL[1]

Abstract. Combinatorial optimization problems such as finding submatrices with large average value within a large data matrix arise in a wide array of fields, ranging from statistical genetics, bioinformatics, computer science to various social sciences. These techniques play an important role in revealing substructures and associations with interesting characteristics in high dimensional problems. In this paper we analyze asymptotics for such problems in an idealized setting where the underlying matrix is a large Gaussian random matrix and provide detailed asymptotics for various characteristics of the energy landscape for such problems. For fixed k we provide a structure theorem for the $k \times k$ submatrix with the largest average. We then show that for any given $\tau > 0$, the size of the largest square sub-matrix with average bigger than τ satisfies a two point concentration phenomena. Finding such submatrices for a fixed k is a computationally intensive problem. We study the natural algorithm that attempts to find submatrices with large average; such algorithms typically converge to a local optimum. We prove a structure theorem for such locally optimal sub-matrices and derive refined asymptotics for the mean and the variance for $L_n(k) :=$ number of such local optima. In particular for $k = 2$ and $k = 3$, the order of the means are n^2 and n^3, while the variances are $n^{8/3}$ and $n^{9/2}$, respectively, with logarithmic corrections. We develop a new variant of Stein's method to prove a Gaussian Central Limit Theorem for $L_n(k)$ for all finite k.

1. Introduction

The study of random matrices, at the level of the empirical spectral distribution, has been one of the fundamental thrusts of modern probability theory. In the last few years, motivated by the explosion in data generated in biology especially genetics [24], as well as complex networks [17], the problem of finding interesting structures or patterns within a large data matrix has become an important research direction. The simplest example of such structures are submatrices with large average. In the context of genetics, such submatrices represent interesting patient to gene relationships, useful as a first step in identifying genes relevant to a disease (see [31] and the references therein). In the context of networks with matrices representing the strength of interaction between different individuals in the network, such submatrices (especially with the same row and column set) are thought to represent "communities" within the network. Finding submatrices with large average or

[1] Department of Statistics and Operations Research, 304 Hanes Hall, University of North Carolina, Chapel Hill, NC 27599

[2] Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012

E-mail addresses: bhamidi@email.unc.edu, partha@cims.nyu.edu, nobel@email.unc.edu.

2010 *Mathematics Subject Classification.* Primary: 62G32, 60F05, 60G70.

Key words and phrases. Energy landscape, Extreme value theory, Central limit theorem, Stein's method.

average above a certain threshold plays a crucial role in various exploratory analysis in such situations.

Now suppose we are given a $n \times n$ square matrix $\mathbf{W}$. Write $[n] := \{1, 2, \ldots, n\}$ for the row set (alternatively the column set) of $\mathbf{W}$. Fix $k \geqslant 2$ and consider the task of finding the $k \times k$ submatrix amongst all possible $\binom{n}{k}^2$ such submatrices with the largest average. For small k one can conceivably tabulate the average of all such submatrices, however the configuration space grows very quickly as k increases and such complete enumeration strategies start to become infeasible even for moderately large k in the context of the scale of data one has in practice.

A number of iterative algorithms have been proposed to find such submatrices. One of the simplest such algorithms, often referred to as **LAS** (Large average submatrix, [31]) operates as follows. Represent a submatrix λ as $\lambda := A \times B$ where $A, B \subseteq [n]$ represent the row and column set respectively of the submatrix. Start with a randomly chosen set A_0 of k rows, find the k columns, say B_0 with largest sums. Set $\lambda_0 = A_0 \times B_0$. Call this the **column step**; intuitively for a fixed set of rows we have found the "best" columns. Now proceed to the **row step**, where we keep the set of columns, namely B_0, fixed and find the set of rows with the largest row sums say A_1 and now let $\lambda_1 = A_1 \times B_0$. These two steps are iterated until one converges. This happens when the algorithm reaches a submatrix $\lambda^* = A^* \times B^*$ which is a **local optimum**, namely the minimal row sum of the submatrix λ^* is larger than the maximal row sum of $([n] \setminus A^*) \times B^*$ and the minimal column sum of λ^* is larger than the maximal column sum of $A^* \times ([n] \setminus B^*)$. Empirical findings in [31] seemed to suggest that both in the context of empirical data as well as simulated data, this algorithm converged quickly and for real data, found matrices with interesting and interpretable structure.

These empirical findings motivated us to provide a rigorous understanding of the general "energy" landscape of such problems in the simplest idealized setting where the underlying matrix $\mathbf{W} = ((w_{ij}))_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n}$ is a gaussian random matrix. Studying optimization problems and properties of optimal or locally optimal configurations for random data has now blossomed into a thriving subbranch of probability, see e.g. [32], [3] and particularly relevant to the kinds of questions addressed in this paper, see [15], [16] and [23]. In our context we are motivated by the following questions:

(i) For a fixed k understand asymptotics for the average of the global optima as well as the structure of the optimal submatrix. Theorem 3.1 gives a description of these asymptotics. On the way we develop a new gaussian comparison result (Lemma 3.2) which is interesting in its own right.

(ii) In the context of applications especially in the genome sciences, for fixed τ, one is interested in finding the largest $k := k(\tau)$ for which there is a $k \times k$ submatrix with average greater than this threshold τ. We prove two point localization for $k(\tau)$ (Theorem 3.4).

(iii) For a fixed k understand asymptotics for the local optima of the **LAS** algorithm. The study of local optima in the context of exploratory data analysis has witnessed renewed interest over the last few years (see e.g. [25]). We give a complete structure theorem for the asymptotic distribution of a locally optimal matrix (Theorem 3.7). A simple corollary of this implies that asymptotically, the average of a typical local optimum is within a factor of $1/\sqrt{2}$ of the global optima.

(iv) We study the number of local optima $L_n(k)$ (Theorem 3.8, 3.9). We derive refined bounds on the means and variances. These results heuristically suggest that the LAS algorithm converges in $\Theta_P((\log n)^{(k-1)/2})$ steps. More interestingly, the variance of $L_n(k)$ has non-standard scaling (for example for $k = 2$, $\mathbb{E}(L_n(2)) \sim n^2$ while $\operatorname{Var}(L_n(2)) \sim n^{8/3}$). The reasons behind these results are the rather mysterious scaling of the positive correlations of pairs of locally optimal submatrices captured in Lemma 3.10 and Lemma 3.11.

(v) We conclude by proving a central limit theorem for $L_n(k)$ (Theorem 3.12). In part due to the above non-standard scaling and highly complex dependency structure, current variants of Stein's method do not seem to apply to this situation and we develop a new variant of Stein's method suitable to this setting.

1.1. **Structure of the paper.** The remaining paper is organized as follows. After a brief summary of notations used in this paper in Section 2, we present the main results in Section 3. We provide more background for the problems studied in this paper and connections between our work to existing literature in Section 4. We dive into proofs in Section 5 which collects some of the technical estimates we need for the proofs of the main results. The reader is urged to skim through these results and then come back to them as and when they are used. We complete the proofs about global optima in Section 6 and local optima in Section 7. Finally we present the proof of the central limit theorem for number of local optimal sub matrices in Section 8.

2. Notation

Given two integers $a \le b$, define $[a, b] := \{a, a+1, \dots, b-1, b\}$. When $a = 1$, instead of $[1, b]$ we will write $[b]$ for simplicity. We will use bold alphabets, e.g. $\mathbf{W}$, for denoting matrices and small alphabets, e.g. w_{ij}, for denoting the corresponding entries.

We shall construct all the finite n problems on the same probability space using a two-dimensional infinite array of i.i.d. standard Gaussian Random variables $\mathbf{W} := ((w_{ij}))_{i,j\ge1}$. For a fixed integer $k \ge 1$, let $\mathscr{S}_n(k)$ be the collection of pairs of subsets of $[n] := \{1, 2, \dots, n\}$ of size k, *i.e.,*

$$\mathscr{S}_n(k) := \{I \times J \mid I, J \subseteq [n], |I| = |J| = k\}.$$

Note that $|\mathscr{S}_n(k)| = \binom{n}{k}^2$. For $\lambda, \gamma \in \mathscr{S}_n(k)$, we write $|\lambda \cap \gamma| = (s, t)$ if λ and γ share s many rows and t many columns. We will also write $\lambda \cap \gamma = \emptyset$ if $|\lambda \cap \gamma| = (s, t)$ with $st = 0$. In this case λ, γ are disjoint, they share no entries.

For $\lambda = I \times J \in \mathscr{S}_n(k)$, define $\mathbf{W}_\lambda$ as the sub-matrix $((w_{ij}))_{i\in I, j\in J}$. For any matrix $\mathbf{U} = ((u_{ij}))$, we define

$$\operatorname{avg}(\mathbf{U}) = |\mathbf{U}|^{-1} \sum u_{ij}$$

as the average of the entries of the sub matrix $\mathbf{U}$, here $|\mathbf{U}|$ is the number of entries of $\mathbf{U}$.

Define

$$\Phi(x) := \mathbb{P}(Z \le x) \text{ and } \bar{\Phi}(x) := 1 - \Phi(x) \text{ for } x \in \mathbb{R}$$

where Z is a standard Gaussian random variable. We shall use $\binom{n}{x}$ to denote the usual binomial coefficients and shall extend the definition for non-integer valued arguments using the Gamma function, in particular for any $x \in [1, n]$, define

$$\binom{n}{x} := \frac{n!}{\Gamma(x+1)\Gamma(n-x+1)} \tag{2.1}$$

where $\Gamma(\alpha) := \int_0^\infty x^{\alpha-1}e^{-x}\,dx$ is the Gamma function. Note that for an integer $k \geq 0$, $\Gamma(k+1) = k!$. Define the sequences

$$a_N := \sqrt{2\log N} \tag{2.2}$$

and

$$b_N := \sqrt{2\log N} - \frac{\log(4\pi\log N)}{2\sqrt{2\log N}}. \tag{2.3}$$

These will arise respectively as scaling and centering constants in the statement of some of our results. Finally given any matrix $\mathbf{U} = ((u_{ij}))$, we shall let $u_{i.}$ denote the average of row i, $u_{.j}$ denote the average of column j and $u_{..} = \operatorname{avg}(\mathbf{W})$. We will write $\tilde{\mathbf{U}} = ((\tilde{u}_{ij}))$ for the the Analysis of variance (ANOVA) decomposition of the matrix $\mathbf{W} = ((w_{ij}))$ namely

$$\tilde{u}_{ij} = u_{ij} - u_{i.} - u_{.j} + u_{..} \tag{2.4}$$

3. Main results

We now state the main results in the paper. We start with asymptotics about the global optima in Section 3.1. We then analyze local optima in Section 3.2.

3.1. Global optima.

3.1.1. *Structure of the Globally optimal sub-matrix.* For an integer $k \geq 1$, define

$$\lambda^*(k) := \operatorname{argmax}\{\operatorname{avg}(\mathbf{W}_\lambda) \mid \lambda \in \mathscr{S}_n(k)\}$$
$$M_n(k) := \max\{\operatorname{avg}(\mathbf{W}_\lambda) \mid \lambda \in \mathscr{S}_n(k)\}.$$

In particular, $\lambda^*(k)$ is the row-column index for the globally optimal sub-matrix and $M_n(k)$ is the average of that matrix. We will prove a complete structure theorem for $\mathbf{W}_{\lambda^*(k)}$ in Theorem 3.1. Recall that for a fixed k there are $N = \binom{n}{k}^2$ many square submatrices of size k. The first part of Theorem 3.1 says that the global optimal average has the same distributional asymptotics as that of the maximum of N i.i.d. Gaussian random variables as long as $k := k(n) \leqslant c\log n/\log\log n$. We believe the result should be true as long as $k \ll \log n$, however this extension will require new ideas. The second part of the theorem says that to first order, such first order asymptotics continue to hold as long as $\log k \ll \log n$. The third part gives a structure theorem for the actual matrix $\lambda^*(k)$ for fixed k.

Theorem 3.1. *Let $N = \binom{n}{k}^2$ and recall the constants a_N and b_N from (2.2) and (2.3).*

(a) *There exists a constant $c > 0$ such that for $k \leq c\log n/\log\log n$ we have*

$$a_N(kM_n(k) - b_N) \overset{\mathrm{d}}{\Longrightarrow} -\log T$$

as $n \to \infty$ where $T \sim \mathit{Exp}(1)$.

(b) *In general, for $c\log n/\log\log n \leq k \leq \exp(o(\log n))$ we have*

$$\mathbb{P}(-k\omega_n(\log\log n)^2/\log n) \leq a_N(kM_n(k) - b_N) \leq \omega_n) \to 1$$

for any function $\omega_n \to \infty$ as $n \to \infty$.

(c) *Moreover, for each fixed integer $k \geqslant 1$ we have*

$$\mathbf{W}_{\lambda^*(k)} - \operatorname{avg}(\mathbf{W}_{\lambda^*(k)})\mathbf{1}\mathbf{1}' \overset{\mathrm{d}}{\Longrightarrow} \mathbf{W}_{[k]\times[k]} - \operatorname{avg}(\mathbf{W}_{[k]\times[k]})\mathbf{1}\mathbf{1}'$$

where $\mathbf{1}$ is the column vector of all ones.

Our main ingredient will be the following comparison Lemma, which is of independent interest.

Lemma 3.2. *Fix $N \geqslant 2$ and let $(X_1, X_2, \ldots, X_N)$ be jointly Gaussian with $\mathbb{E}(X_i) = 0, \mathbb{E}(X_i^2) = 1$ and $\mathbb{E}(X_iX_j) = \sigma_{ij} \in (-1,1)$ for $1 \le i < j \le N$. Let $Z_1, Z_2, \ldots, Z_N$ be i.i.d. standard Gaussian random varaibles. For any $u \ge 1$, we have*

$$\begin{aligned}
&|\,\mathbb{P}(\max_{1\le i\le N} X_i \le u) - \mathbb{P}(\max_{1\le i\le N} Z_i \le u)| \\
&\quad\le \sum_{1\le i<j\le N} 2\min\{1, |1-\theta_{ij}|u(u+1)\}\bar\Phi(u)\bar\Phi((1\wedge\theta_{ij})u) \\
&\quad\le \sum_{i<j,\sigma_{ij}\neq 0} 2\sqrt{\frac{1+\sigma_{ij}^+}{1-\sigma_{ij}^+}}\cdot\bar\Phi(u)^2\cdot e^{u^2\sigma_{ij}^+/(1+\sigma_{ij}^+)}
\end{aligned}$$

where $\theta_{ij} = \sqrt{(1-\sigma_{ij})/(1+\sigma_{ij})}$ and $x^+ = \max\{x, 0\}$.

Note that as a corollary, when $\sigma_{ij} \ge 0$ for all i, j, we have

$$0 \le \mathbb{P}(\max_{1\le i\le N} X_i \le u) - \mathbb{P}(\max_{1\le i\le N} Z_i \le u) \le \mathbb{E}(\mathcal{N}(u)^2) - N\bar\Phi(u)^2$$

where $\mathcal{N}(u) = \sum_{i=1}^N \mathbb{1}\{X_i \ge u\}$ and the first inequality is by Slepian's lemma. Various gaussian comparison results similar to Lemma 3.2 are known (see [7, 18, 22]). This variant seems best suited for our purposes and in particular allows us to extend first order asymptotics for $k = k(n) \to \infty$ as in Theorem 3.1 (b).

3.1.2. *Two-point localization.* Fix $\tau > 0$. Let $M^* = M^*(\tau)$ denote the largest k such that there exists a $k \times k$ sub matrix $\mathbf{W}_\lambda$ with $\lambda \in \mathscr{S}_n(k)$ and $\operatorname{avg}(\mathbf{W}_\lambda) > \tau$. The next theorem states that for each fixed $\tau > 0$, the random variable $M^*(\tau)$ localizes on at most two consecutive values as $n \to \infty$.

Recall the definition of the binomial coefficient for non-integer values from (2.1). Let $\tilde k = \tilde k_n(\tau) > 1$ denote the unique solution (for large n) of the equation

$$\binom{n}{\tilde k}^2 \bar\Phi(\tilde k\tau) := 1. \tag{3.1}$$

It is easy to check (see e.g. [34]) that

$$\tilde k = \frac{4}{\tau^2}\log\frac{e\tau^2 n}{4\log n} + \left(\frac{4}{\tau^2} - 1\right)\frac{\log\log n}{\log n} + O\left(\frac{|\log\tau|}{\tau^2\log n}\right) \tag{3.2}$$

as $n \to \infty$. Let k^* denote the closest integer to $\tilde k$. In [34] it was proved that

Theorem 3.3 (Theorem 1 in [34])**.** *For fixed $\tau > 0$, we have*

$$\mathbb{P}\left(-\frac{4}{\tau^2} - \frac{12\log 2}{\tau^2} - 4 \le M^*(\tau) - \tilde k \le 2\right) \to 1$$

as $n \to \infty$.

We improve the result to a two point localization result.

Theorem 3.4 (Localization for fixed threshold τ)**.** *We have*

$$\mathbb{P}(M^*(\tau) = k^* \text{ or } k^* - 1) \to 1$$

as $n \to \infty$.

3.2. **Locally optimal sub matrices.** Fix $k \geq 1$ and recall the algorithm described in Section 1 designed to detect $k \times k$ sub-matrices with large average. Note that this algorithm terminates at a local optima. By definition a submatrix is a local optima if it is an optima in the column step **and** row step of the algorithm. The next few results give asymptotics for the distribution of a typical local optima as well as the number of such local optima. For future reference we first formalize these definitions.

Definition 3.5. *Given two subsets $\lambda = I \times J \in \mathscr{S}_n(k)$, we call the sub-matrix $\mathbf{W}_\lambda := ((W_{ij}))_{i \in I, j \in J}$ row optimal if*

$$\operatorname{avg}(\mathbf{W}_{I\times J}) = \max_{|J'|=k} \operatorname{avg}(\mathbf{W}_{I\times J'})$$

and column optimal if

$$\operatorname{avg}(\mathbf{W}_{I\times J}) = \max_{|I'|=k} \operatorname{avg}(\mathbf{W}_{I'\times J})$$

A submatrix which is both row and column optimal is called locally optimal.

We are interested in the distribution of $\mathbf{W}_\lambda$ conditioned on being a local optima (by symmetry the choice of λ is irrelevant) as well as the number of such local optima namely,

$$L_n(k) := \sum_{\lambda \in \mathscr{S}_n(k)} \mathbb{1}\{\mathbf{W}_\lambda \text{ is locally optimal}\}, \tag{3.3}$$

the number of $k \times k$ locally optimal submatrices. Note that, for any fixed set of k rows $I \subseteq [n]$, there is a unique row optimal sub matrix $\mathbf{W}_{I \times J^*(I)}$. Thus we can also write

$$L_n(k) = \sum_{|I|=k} \mathbb{1}\{\mathbf{W}_{I\times J^*(I)} \text{ is column optimal }\}. \tag{3.4}$$

From (3.4) it is easy to see that

$$\mathbb{E}(L_n(k)) = \binom{n}{k} \mathbb{P}(\mathbf{W}_{\lambda_k} \text{ is column optimal} \mid \mathbf{W}_{\lambda_k} \text{ is row optimal})$$

where $\lambda_k = [k] \times [k] \in \mathscr{S}_n(k)$. Thus our first order of business is understanding the asymptotic conditional probability of a submatrix being a column optimal given it is row optimal; by symmetry this is the same for all $k \times k$ submatrices. Intuitively one might expect $\mathbb{P}(\mathbf{W}_{\lambda_k}$ is column optimal $\mid \mathbf{W}_{\lambda_k}$ is row optimal$)= O(1)$ as $n \to \infty$ since in some sense we have already conditioned the entries of $\mathbf{W}_{\lambda_k}$ to be large. Indeed for $k = 1$, we have $\mathbb{P}(\mathbf{W}_{\lambda_1}$ is column optimal $\mid \mathbf{W}_{\lambda_1}$ is row optimal $) = n/(2n-1) \approx 1/2$. However, it turns out that, for $k \geq 2$, the conditional probability that a row-optimal matrix is also column optimal vanishes as $n \to \infty$ and in fact behaves like $(\log n)^{-(k-1)/2}$. We first give some intuition behind this phenomenon and then state a precise structure theorem for the distribution of a local optima. We will need the following standard result from extreme value theory, see e.g. [21].

Lemma 3.6. *Let $Z_1, Z_2, \ldots, Z_n$ be n i.i.d. standard Gaussian random variables and $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)}$ be their ordered values. Recall the scaling and centering constants a_n, b_n from (2.2) and (2.3). Then for any fixed integer $\ell \geq 0$, we have*

$$a_n(Z_{(n)} - b_n, Z_{(n-1)} - b_n, \ldots, Z_{(n-\ell)} - b_n) \Rightarrow (V_1, V_2, \ldots, V_\ell)$$

as $n \to \infty$ where $V_i = -\log(T_1 + T_2 + \cdots + T_i), i \geq 1$ and T_i's are i.i.d. Exponential(1) random variables. Moreover, the point process $\sum_{i=1}^{n} \delta_{a_n(Z_i - b_n)}$ converges as $n \to \infty$ to a Poisson Point Process on $\mathbb{R}$ with intensity e^{-x}.

Using Lemma 3.6 one can see that all the column averages of a row optimal matrix $\mathbf{U}$ (and hence the matrix average $\mathrm{avg}(\mathbf{U})$) will be concentrated around $k^{-1/2}b_n$ with $O(1/a_n)$ fluctuation. Recall that $a_n = \sqrt{2\log n}$. Now for a Gaussian random matrix with i.i.d. entries the centered row averages are independent of the matrix average. Hence the minimum row average will be the same as the matrix average $\mathrm{avg}(\mathbf{U}) - \ V$, where $V \stackrel{\mathrm{d}}{=} k^{-1/2} \max_{1 \leq i \leq k}(\bar{Z} - Z_i)$. Here $\{Z_i\}_{1 \leqslant i \leqslant k}$ are i.i.d. standard Gaussian r.v.s. Note that we have fixed $\lambda_k = [k] \times [k]$. Fixing this column set, the maximum of all other row averages (amongst rows $[n] \setminus [k]$), by Lemma 3.6, will be concentrated around $k^{-1/2}b_n$ with $O(1/a_n)$ fluctuations. In order for the row optimal matrix to now be column optimal, one needs V to be of the order of $1/a_n$, the probability of which turns out to be of order $1/a_n^{k-1}$ as the vector $(\bar{Z} - Z_1, \bar{Z} - Z_2, \ldots, \bar{Z} - Z_k)$ lies in a $(k-1)$-dimensional subspace.

Let $\mathcal{I}_k$ be the event that $\mathbf{W}_{[k]\times[k]}$ is locally optimal. We will prove the following structure theorem.

Theorem 3.7 (Structure Theorem for locally optimal submatrices)**.** *Conditional on the event $\mathcal{I}_k$, the sub matrix $\mathbf{W}_{[k]\times[k]}$ can be written as*

$$\mathbf{W}_{[k]\times[k]} \stackrel{\mathrm{d}}{=} \left(\frac{b_n}{\sqrt{k}} - \frac{\log G}{\sqrt{k}a_n}\right)\mathbf{1}\mathbf{1}' + \tilde{\mathbf{Z}}$$

$$+ \frac{\log(1 + T/G)}{\sqrt{k}a_n}\begin{bmatrix} kU_1 - 1 \\ kU_2 - 1 \\ \vdots \\ kU_k - 1 \end{bmatrix}\mathbf{1}' + \frac{\log(1 + T'/G)}{\sqrt{k}a_n}\mathbf{1}\begin{bmatrix} kU_1' - 1 \\ kU_2' - 1 \\ \vdots \\ kU_k' - 1 \end{bmatrix}' + o_p(1/a_n)$$

where $\tilde{\mathbf{Z}} = (\tilde{z}_{ij})_{k\times k}$ with $\tilde{z}_{ij} = z_{ij} - z_{i\cdot} - z_{\cdot j} + z_{\cdot\cdot}$, z_{ij}'s are i.i.d. standard Gaussian, $\mathbf{U} = (U_1, U_2, \ldots, U_k)$, $\mathbf{U}' = (U_1', U_2', \ldots, U_k')$ are i.i.d. from Dirichlet$(1, 1, \ldots, 1)$ distribution independent of (G, T, T') which have joint density

$$\propto (\log(1 + t/g)\log(1 + t'/g))^{k-1} g^{k-1} e^{-t-t'-2g}, \quad g, t, t' \geq 0.$$

Further, there exists a real number $\theta_k > 0$ such that we have

$$\mathbb{P}(\mathcal{I}_k) = \frac{\theta_k}{\binom{n}{k}(\log n)^{(k-1)/2}}(1 + o(1)) \ \textit{as } n \to \infty.$$

The value of θ_k can be explicitly computed as

$$\theta_k := \frac{k^{2k+1/2}}{2^{2k-1}\pi^{(k-1)/2}k!^2}\,\mathbb{E}((\log(1 + Y/G)\log(1 + Y'/G))^{k-1}) \tag{3.5}$$

where Y, Y' are i.i.d. $\mathrm{Exp}(1)$ and $G \sim \mathrm{Gamma}(k, 2)$ with density $\frac{2^k}{(k-1)!}x^{k-1}e^{-2x}, x > 0$ independent of Y, Y' (see equation (7.6)).

Note that as a corollary we have, all the entries in a typical locally optimal sub matrix are concentrated at $\sqrt{2\log n/k}(1 + o(1)$. By symmetry, the expectation of the number of local maxima $L_n(k)$ is $\mathbb{E}(L_n(k)) = \binom{n}{k}^2 \mathbb{P}(\mathcal{I}_k)$. Thus theorem 3.7 immediately yields the following result.

Theorem 3.8 (Mean behavior)**.** *For any fixed $k \geq 1$, the expected number of local maxima scales like*

$$\mathbb{E}(L_n(k)) = \frac{\theta_k \binom{n}{k}}{(\log n)^{(k-1)/2}}(1 + o(1)) \text{ as } n \to \infty.$$

Here $\theta_k > 0$ is as in (3.5).

Intuitively this suggests that the running time of the **LAS** algorithm can be bounded by a Geometric random variable with $p = p(n) = \theta_k/(\log n)^{(k-1)/2}$, and thus converges in $\Theta_P((\log n)^{(k-1)/2})$ steps, and thus gives conceptual insight on empirical observations on the running time of the algorithm.

The behavior of the variance of $L_n(k)$ is much more delicate and requires analyzing the joint distribution of two locally optimal sub-matrices and in particular yields non-standard scaling as described in the Introduction, in particular $\mathrm{Var}(L_n(k)) >> \mathbb{E}(L_n(k))$. The reason behind the high value for the variance is the complex dependence structure amongst locally optimal matrices.

Theorem 3.9 (Variance behavior)**.** *There exists $\nu_k \in (0, \infty)$ such that*

$$\mathrm{Var}(L_n(k)) = \frac{\nu_k n^{2k^2/(k+1)}}{(\log n)^{k^2/(k+1)}}(1 + o(1)) \text{ as } n \to \infty.$$

There are two main ingredients in this variance calculation which are of independent interest and could conceivably be of use in the analysis of similar iterative methods. The first is the following lemma which gives the probability that both the maximum row average and maximum column average of a Gaussian random matrix with i.i.d. entries are large.

Lemma 3.10. *Let $\mathbf{U}$ be a $s \times t$ matrix of i.i.d. standard Gaussian entries. Recall that $u_{i\cdot}, u_{\cdot j}$ was respectively the i-th row average and j-th column average of $\mathbf{U}$. For any fixed $\theta > 0, x, y \in \mathbb{R}$, we have*

$$\begin{aligned}
&\mathbb{P}(\max_{1 \leq i \leq s} u_{i\cdot} \geq \theta b_n + x/a_n, \max_{1 \leq j \leq t} u_{\cdot j} \geq \theta b_n + y/a_n) \\
&= (\eta(s,t,\theta) + o(1)) e^{-st\theta((t-1)x+(s-1)y)/(st-1)} n^{-\frac{st(s+t-2)\theta^2}{st-1}} (\log n)^{\frac{st(s+t-2)\theta^2}{2(st-1)} - 1}
\end{aligned}$$

for some constant $\eta(s,t,\theta) > 0$.

The heuristic idea behind the proof of Lemma 3.10 is the following. If both the maximum row average and maximum column average are bigger than z, there will be at least one row (say i_*-th row) and one column (say j_*-th column) with average bigger that z. The joint density of the i_*-th row and j_*-th column is proportional to $\exp(-(\sum_{i \neq i_*} u_{ij_*}^2 + \sum_{j \neq j_*} u_{i_*j}^2 + u_{i_*j_*}^2)/2)$. If we minimize $\sum_{i \neq i_*} u_{ij_*}^2 + \sum_{j \neq j_*} u_{i_*j}^2 + u_{i_*j_*}^2$ under the constraint that $\sum_i u_{ij_*} \geq tz, \sum_j u_{i_*j} \geq sz$, the minimum is achieved at

$$u_{ij_*} = \frac{(st-s)z}{st-1} \text{ for } i \neq i_*, \qquad u_{ij_*} = \frac{(st-t)z}{st-1} \text{ for } j \neq j_*$$
$$\text{and } u_{i_*j_*} = \frac{(2st-s-t)z}{st-1}.$$

Plugging in these values in the exponent results in the value $st(s+t-2)z^2/(st-1)$. When $z = \theta b_n$, we have

$$\exp(-\frac{st(s+t-2)z^2}{2(st-1)}) \approx n^{-\frac{st(s+t-2)\theta^2}{st-1}},$$

which is the leading order in the probability. The complete proof is given in Section 5.

The other ingredient is the following joint probability estimate for two locally optimal matrices. We will need to setup some notation. Fix two integers $s, t \in [k]$. Let $\mathcal{B}_{s,t,k}$ be the event that $\mathbf{W}_{[k]\times[k]}$ is locally optimal as a sub matrix of $\mathbf{W}_{[k]\cup[s+k+1,n]\times[k]\cup[t+k+1,n]}$ and $\mathbf{W}_{[s+1,s+k]\times[t+1,t+k]}$ is locally optimal as a sub matrix of $\mathbf{W}_{[s+1,n]\times[t+1,n]}$ (see Figure 3.1).

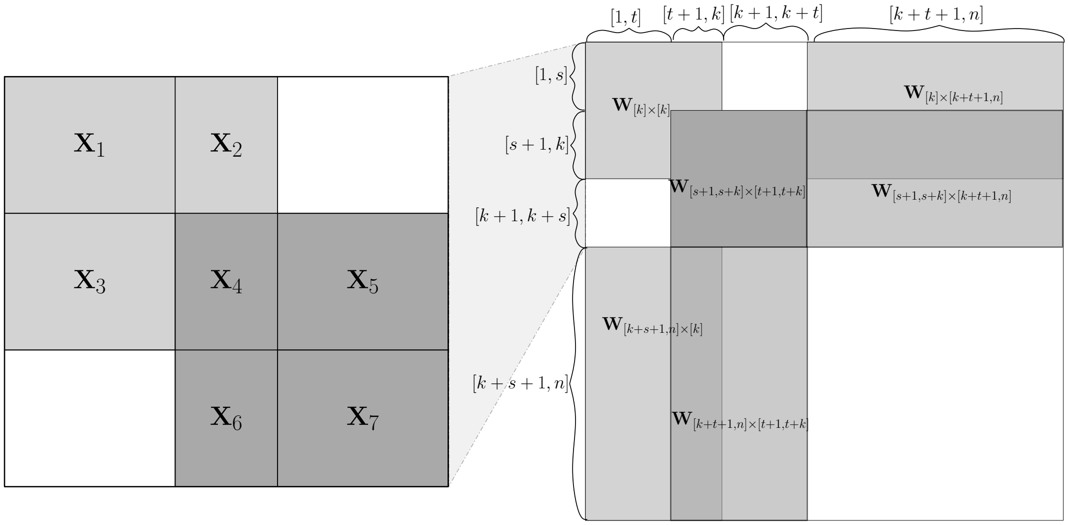
$\mathbf{X}_1$
$\mathbf{X}_2$
$\mathbf{X}_3$
$\mathbf{X}_4$
$\mathbf{X}_5$
$\mathbf{X}_6$
$\mathbf{X}_7$
$[1,t]$
$[t+1,k]$
$[k+1,k+t]$
$[k+t+1,n]$
$[1,s]$
$[s+1,k]$
$[k+1,k+s]$
$[k+s+1,n]$
$\mathbf{W}_{[k]\times[k]}$
$\mathbf{W}_{[k]\times[k+t+1,n]}$
$\mathbf{W}_{[s+1,s+k]\times[t+1,t+k]}$
$\mathbf{W}_{[s+1,s+k]\times[k+t+1,n]}$
$\mathbf{W}_{[k+s+1,n]\times[k]}$
$\mathbf{W}_{[k+t+1,n]\times[t+1,t+k]}$

FIGURE 3.1. A pictorial representation of the event $\mathcal{B}_{s,t,k}$ and the block matrices $\mathbf{X}_i, 1 \le i \le 7$.

Lemma 3.11. *Let $0 < s < k, 0 < t < k$. There exists a constant $\eta(s,t,k) > 0$ such that*

$$\mathbb{P}(\mathcal{B}_{s,t,k}) \le \eta(s,t,k) \left(\frac{\sqrt{\log n}}{n}\right)^{2k-2k(k-s)(k-t)/(2k^2-st)}.$$

It clearly follows that the joint probability $\mathbb{P}(\mathcal{B}_{s,t,k})$ is much bigger than the product of the probabilities that both $\mathbf{W}_{[k]\times[k]}$ and $\mathbf{W}_{[s+1,s+k]\times[t+1,t+k]}$ are locally optimal. Now note the block matrix decomposition shown in Figure 3.1. Under the event $\mathcal{B}_{s,t,k}$, the average entry in $\mathbf{X}_4, \mathbf{X}_1, \mathbf{X}_7$ are much larger than the average entries in $\mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_5, \mathbf{X}_6$. However the global average is still close to $\sqrt{2\log n/k}$ which is the average of a typical locally optimal matrix.

Finally, using Stein's method we will prove asymptotic normality for the random variable $L_n(k)$. Despite being able to express L_n as a sum of random variables

$$L_n(k) := \sum_{\lambda\in\mathscr{S}_n(k)} \mathbb{1}\left\{\mathbf{W}_\lambda \text{ is locally optimal in } \mathbf{W}_{[n]\times[n]}\right\},$$

as opposed to typical settings of weak dependence, here the event that a particular submatrix $\mathbf{W}_\lambda$ is locally optimal affects **every** other submatrix $\lambda' \in \mathscr{S}_n(k)$. The analysis of $\mathrm{Var}(L_n(k))$ suggests non-trivial correlations. This makes the analysis of the asymptotic distribution particularly involved.

Recall that the Wasserstein distance between two random variables W, Z is defined as

$$d_{\mathcal{W}}(W,Z) := \sup\left\{|\mathbb{E}(g(W)) - \mathbb{E}(g(Z))| : g(\cdot)\ 1-\text{Lipschitz}\right\}$$

The following result quantifies the distance from normality of $L_n(k)$.

Theorem 3.12 (Central Limit Theorem for $L_n(k)$)**.** *We have,*

$$\tilde{L}_n(k) := \frac{L_n(k) - \mathbb{E}(L_n(k))}{\sqrt{\operatorname{Var}(L_n(k))}} \overset{\mathrm{d}}{\Longrightarrow} N(0,1)$$

as $n \to \infty$. Moreover, we have

$$d_{\mathcal{W}}(\tilde{L}_n(k), N(0,1)) \le n^{-\frac{2(k-1)}{(k+1)(k^2+2k-1)}+O(\log\log n/\log n)}.$$

where $d_{\mathcal{W}}$ is the Wasserstein distance between two distributions.

Here we mention that the rate of convergence in Theorem 3.12 is definitely not optimal and we haven't tried to find the optimal rate. However, we include simulation results in Figure 3.2 for $k = 2$ and $n \in \{100, 200\}$ with 5000 runs to show the fast convergence to Gaussian limit empirically.

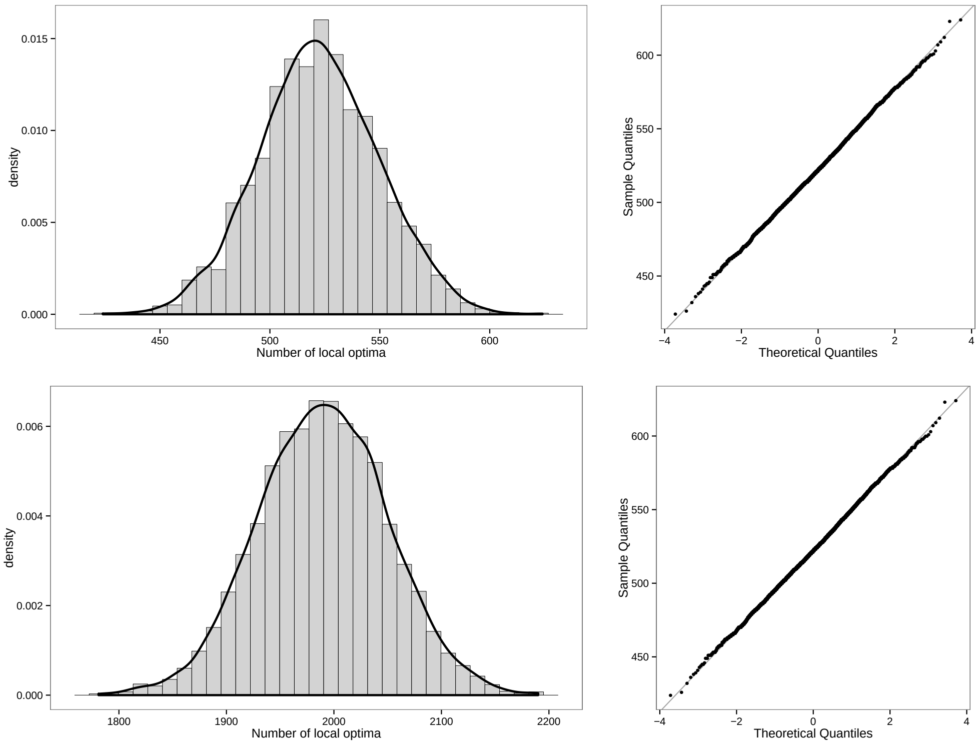

FIGURE 3.2. Histogram and QQPlot for number of local optima for $k = 2$ with $n = 100$ (top row) and $n = 200$ (bottom row) with 5000 samples.

4. DISCUSSION

We now discuss the relevance of these results and related work. We start with a discussion of the general detection problem considered in this work and then expand on the techniques used in the paper.

4.1. **Finding large substructures.** As mentioned above, with the advent of large scale data in genomics, problems such as finding interesting structures in matrices has stimulated a lot of interest in a number of different communities, see e.g. the survey [24]. In spirit, such problems are linked to another large body of work in the combinatorics community, namely the hidden clique problem see e.g. [28] or [19] and the references therein. The simplest statement of the problem is as follows: Select a graph at random on n vertices; consider the problem of detecting the largest clique (fully connected subgraph). For large n, it is known that the largest clique has $k(n) \sim 2\log_2 n$ vertices ([8], [9]). Theorem 3.4 is very similar, in spirit to this result. However most greedy heuristics and formulated algorithms, short of complete enumeration, are only able to find cliques of size $\sim \log_2 n$ and thus are off by a factor of 2 from the optimal size. We see analogous behavior in our results; Theorem 3.1(a) implies that for fixed k, the average of the global optimum scales like $\sqrt{2}\sqrt{\log n/k}$ whilst Theorem 3.7 implies that the average of a typical local optima scales like $\sqrt{\log n/k}$.

4.2. **Planted detection problems.** In the context of statistical testing of hypothesis, we have analyzed the energy landscape in the "null" case. One could also look at the "alternative" where there is some inherent structure in the data. In the last few years there has been a lot of interest in formulating statistical tests of hypothesis to distinguish between the null and the alternative, see e.g. [5] and [6] for the general framework as well as application areas motivating such questions and see [1] and [10] for a number of interesting general results in these contexts. In the context of the combinatorics, such questions result in the famous planted clique problem see e.g [4], [13] and the references therein.

4.3. **Energy landscapes.** The notion of energy or fitness landscapes, incorporating a fitness or score to each element in a configuration and then exploring the ruggedness of the subsequent landscape, arose in evolutionary biology, see [37], and for a nice survey, see [29]. Our work has been partially inspired by the rigorous analysis of the NK fitness model ([20], [35]) carried out in the probability community in papers such as [15], [16], [23]. These questions have also played a major role in understanding deep underlying structures in spin glass in statistical physics, see e.g. [27]. For general modern accounts of the state of the art on combinatorial optimization in the context of random data and connections to other phenomenon in statistical physics, we refer the interested reader to [26].

4.4. **Stein's method for normal approximations.** Stein's method is a general and powerful method for proving distributional convergence with explicit rate of convergence. Developed by Charles Stein in 1972 [33], to prove Gaussian central limit theorems for sums of random variables with complex dependency structure. This has now been extended to a wide array of distributions. Here we briefly discuss the case of normal approximation.

The standard Gaussian distribution can be characterized by the operator $\mathcal{A}f(x) := xf(x) - f'(x)$ in the sense that, X has standard Gaussian distribution iff $\mathbb{E}(\mathcal{A}f(X)) = 0$ for all absolutely continuous functions f. Now to measure the closeness between a distribution ν and the standard Gaussian distribution ν_0, one generally uses a separating class of functions $\mathcal{D}$ to define a distance

$$d_{\mathcal{D}}(\nu, \nu_0) = \sup_{h\in\mathcal{D}} |\,\mathbb{E}\,h(X) - \mathbb{E}\,h(Z)|$$

where $X \sim \nu, Z \sim \mathrm{N}(0,1)$ and then attempts to show that the distance is "small". In this paper we will consider the L^1-Wasserstein distance in which case $\mathcal{D}$ is the class of all 1-Lipschitz functions.

Stein's method consists of two main steps. The first step is to find solution to the equation $\mathcal{A}f_h(x) = h(x) - \mathbb{E}\, h(Z)$ for $h \in \mathcal{D}$. Assuming this can be performed, we have,

$$\sup_{h\in\mathcal{D}} |\,\mathbb{E}\, h(X) - \mathbb{E}\, h(Z)| \leq \sup_{f\in\mathcal{D}'} |\,\mathbb{E}(Xf(X) - f'(X))|$$

where $\mathcal{D}' = \{f_h \mid h \in \mathcal{D}\}$. The following lemma summarizes the the bounds required for Stein's method.

Lemma 4.1 ([33])**.** *For any 1-Lipschitz function h, there is a unique function f_h such that $\mathcal{A}f_h = h - \mathbb{E}\, h(Z)$. Moreover we have*

$$|f_h|_\infty \leq 1, |f_h'|_\infty \leq \sqrt{2/\pi} \text{ and } |f_h''|_\infty \leq 2.$$

Thus to prove that the distribution of X is close to standard Gaussian distribution it is enough to prove that

$$\sup_{f\in\mathcal{D}'} |\,\mathbb{E}\, f'(X) - \mathbb{E}\, Xf(X)|$$

is small where

$$\mathcal{D}' = \{f \mid |f_h|_\infty \leq 1, |f_h'|_\infty \leq \sqrt{2/\pi} \text{ and } |f_h''|_\infty \leq 2\}. \tag{4.1}$$

This final portion is very much problem dependent and is often the hardest to accomplish. A number of general techniques have now been formulated, *e.g.,* exchangeable pair approach, dependency graph approach, size-bias transform, zero-bias transform etc. that can be used for a large class of problems. We refer the interested reader to the surveys [11, 12, 14, 30] and the references therein. However, in our case because of the high degree of dependency, the above mentioned methods are difficult to apply and we develop a new variant to bound the error.

4.5. **Open questions.** For the sake of mathematical tractability, we assumed that the underlying matrix had gaussian entries. It would be interesting to extend this analysis to general distributions. The exact statement of the results will be different since extremal properties of the gaussian distribution play a significant role in the proofs of the main results. The results in the paper also suggest a host of extensions and new problems. Theorem 3.1 deals with the global optimum in the regime where $\log k = o(\log n)$. Extending this further, especially to the regime where $k = \alpha n$ for some $0 < \alpha < 1$ would be quite interesting and will require new ideas; one expects that the comparison to the independence regime using Lemma 3.2 breaks down at this stage. We also expect behavior similar to the extrema of branching random walk ([2] and references within) in this regime. Extending the local optima results to a regime $k = k(n) \to \infty$ as opposed to the fixed k regime would be interesting. This would be especially relevant in the context of detecting matrices with average above a particular threshold which by Theorem 3.4 corresponds to the $k(n) = C \log n$ regime. Finally this work fixes k and then tries to find submatrices with large average. It would be interesting to develop algorithms which allow one to increase k to achieve large submatrices with average above a threshold τ.

5. Technical Estimates

We start with some technical estimates that will be needed in the later proofs. We start with various estimates on the tails of the normal distribution which will then lead to a proof of Lemma 3.10 in Section 5.3. We conclude this section with some combinatorial estimates.

5.1. **Gaussian tail bounds.**

The following is a standard bound on the Gaussian tail, see e.g. [21].

Lemma 5.1. *Let Z be a standard Gaussian r.v. Then we have*

$$\frac{xe^{-x^2/2}}{\sqrt{2\pi}(1+x^2)} \le \mathbb{P}(Z \ge x) \le \frac{e^{-x^2/2}}{\sqrt{2\pi}x} \text{ for } x > 0.$$

The next result uses the above Lemma to understand the conditional distribution of a standard Gaussian conditioned to be large.

Lemma 5.2. *Let Z be a standard Gaussian random variable and $\theta > 0$ be a fixed real number. Let a_n and b_n be as in (2.2) and (2.3). Let $\mathcal{B}_x$ be the event $\{Z \ge \sqrt{\theta}(b_n + a_n^{-1}x)\}$.*

(a) *We have*

$$n^{\theta}(\sqrt{2\pi}b_n)^{1-\theta}e^{x\theta}\,\mathbb{P}(\mathcal{B}_x) \text{ converges to } \theta^{-1/2} \text{ uniformly for } |x| \ll a_n$$

as $n \to \infty$.

(b) *Conditionally on the event $\mathcal{B}_x$, $a_n(Z/\sqrt{\theta} - b_n - a_n^{-1}x)$ converges in distribution to an Exponential rate θ random variable as $n \to \infty$ for all $x \in \mathbb{R}$.*

Proof. (a) We will use Lemma 5.1 and the fact that $n(\sqrt{2\pi}b_n)^{-1}e^{-b_n^2/2} \to 1$ as $n \to \infty$. It clearly follows that

$$\begin{aligned}
&\lim_{n\to\infty} n^{\theta}(\sqrt{2\pi}b_n)^{1-\theta}e^{x\theta}\,\mathbb{P}(\mathcal{B}_x)\\
&= \lim_{n\to\infty} n^{\theta}(\sqrt{2\pi}b_n)^{1-\theta}e^{x\theta}\,\mathbb{P}(Z \ge \sqrt{\theta}(b_n + a_n^{-1}x))\\
&= \lim_{n\to\infty} \frac{1}{\sqrt{2\pi\theta}b_n} n^{\theta}(\sqrt{2\pi}b_n)^{1-\theta}e^{x\theta}\exp(-\theta(b_n + a_n^{-1}x)^2/2) = \theta^{-1/2}.
\end{aligned}$$

(b) Now for the conditional distribution note that

$$\mathbb{P}(a_n(Z/\sqrt{\theta} - b_n - a_n^{-1}x) \ge t \mid \mathcal{B}_x) = \frac{\mathbb{P}(\mathcal{B}_{x+t})}{\mathbb{P}(\mathcal{B}_x)} \to e^{-\theta t}$$

as $n \to \infty$ for every $t \ge 0$. This completes the proof. ■

To analyze asymptotics for the expected number of local optima $\mathbb{E}(L_n(k))$, we will need to understand the distribution of the deviations of a set of Gaussian random variables from the sample mean under various conditioning events. The next Lemma quantifies the results relevant to our treatment.

Lemma 5.3. *Let $Z_1, Z_2, \ldots, Z_k$ be i.i.d. standard Gaussian r.v.s. Let $\bar{Z} = k^{-1}\sum_{i=1}^{k} Z_i$ and $Z_{\min} = \min_{1\le i\le k} Z_i$ be, respectively, the average and minimum of the random vector $(Z_1, Z_2, \ldots, Z_k)$.*

(a) *The random variable* $\bar{Z} - Z_{\min} \geq 0$ *a.s. and we have* $\mathbb{P}(\bar{Z} - Z_{\min} \leq x) = f_k x^{k-1}(1 + o(1))$ *as* $x \downarrow 0$ *where*

$$f_k := \frac{k^{k+1/2}}{k!(2\pi)^{(k-1)/2}}.$$

(b) *Let* $\mathcal{B}_\varepsilon$ *be the event that* $\{\bar{Z} - Z_{\min} \leq \varepsilon\}$ *for* $\varepsilon > 0$. *The conditional distribution of* $\varepsilon^{-1}(\bar{Z} - Z_1, \bar{Z} - Z_2, \ldots, \bar{Z} - Z_k)$ *given* $\mathcal{B}_\varepsilon$ *converges in distribution, as* $\varepsilon \downarrow 0$, *to* $(1 - kU_1, 1 - kU_2, \ldots, 1 - kU_k)$ *where* $\mathbf{U} = (U_1, U_2, \ldots, U_k)$ *follows the Dirichlet distribution with parameter* $(1, 1, \ldots, 1)$, i.e., $\mathbf{U}$ *is uniformly distributed on the simplex* $\{(x_1, x_2, \ldots, x_k) \mid x_1 + x_2 + \cdots + x_k = 1, x_1 \geq 0, \ldots, x_k \geq 0\}$.

(c) *We have*

$$\mathbb{P}(\bar{Z} - Z_{\min} \geq x) = \frac{g_k}{x} \exp\left(-\frac{kx^2}{2(k-1)}\right)(1 + o(1)) \text{ as } x \uparrow \infty$$

for some real number $g_k > 0$.

Proof of Lemma 5.3. (a) The first assertion that $\bar{Z} - Z_{\min} \geq 0$ follows trivially. It is easy to see that $k^{-1}(Z_{\max} - Z_{\min}) \leq \bar{Z} - Z_{\min} \leq Z_{\max} - Z_{\min}$ where $Z_{\max} = \max_{1 \leq i \leq k} Z_i$. In particular,

$$\mathbb{P}(Z_{\max} - Z_{\min} \leq x) \leq \mathbb{P}(\bar{Z} - Z_{\min} \leq x) \leq \mathbb{P}(Z_{\max} - Z_{\min} \leq kx) \text{ for all } x \geq 0.$$

Now one can easily check that

$$\mathbb{P}(Z_{\max} - Z_{\min} \leq x) = \int_{-\infty}^{\infty} k(\Phi(t + x) - \Phi(t))^{k-1}\Phi'(t)dt$$

where $\Phi(t) = \mathbb{P}(Z \leq t)$ is the standard normal distribution function and the right hand side is $\Theta(x^{k-1})$ as $x \downarrow 0$. This proves that $F(x) := \mathbb{P}(\bar{Z} - Z_{\min} \leq x)$ scales like x^{k-1} as $x \downarrow 0$. Now $\bar{Z} - Z_{\min}$ has a density on $\mathbb{R}_+$ and its Laplace transform is given by

$$\begin{aligned} \mathbb{E}(e^{-t(\bar{Z} - Z_{\min})}) &= ke^{-t^2/2k} \int_{\mathbb{R}} e^{-tx}\Phi(x)^{k-1}\Phi'(x)dx \\ &= ke^{(k-1)t^2/2k} \int_{\mathbb{R}} \Phi(x - t)^{k-1}\Phi'(x)dx \text{ for } t \geq 0. \end{aligned} \tag{5.1}$$

The proof follows from the fact that $\mathbb{E}(e^{t\bar{Z}}) = e^{t^2/2k}, \mathbb{E}(e^{tZ_{\min}}) = \int_{\mathbb{R}} ke^{-tx}\Phi(x)^{k-1}\Phi'(x)dx$ and by independence of $\bar{Z}, \bar{Z} - Z_{\min}$ we have $\mathbb{E}(e^{-t(\bar{Z} - Z_{\min})}) = \mathbb{E}(e^{tZ_{\min}})/\mathbb{E}(e^{t\bar{Z}})$. From our previous calculations it thus follows that there is a constant $f_k > 0$ such that $F(x)/f_k x^{k-1} \to 1$ as $x \downarrow 0$. We claim that

$$f_k = \lim_{t \to \infty} \frac{t^{k-1}}{(k-1)!} \mathbb{E}(e^{-t(\bar{Z} - Z_{\min})}).$$

We leave the proof to the interested reader. Using (5.1) and the fact that

$$\Phi(-x) = \frac{1}{\sqrt{2\pi}x} e^{-x^2/2}(1 - O(x^{-2})) \text{ for } x \to \infty$$

we finally have

$$\begin{aligned}
f_k &= \lim_{t\to\infty} \frac{k^2 e^{(k-1)t^2/2k}}{k!(2\pi)^{k/2}} \int_{\mathbb{R}} \left(\frac{t(1+O((t-x)^{-2}))}{t-x}\right)^{k-1} e^{-\frac{1}{2}((k-1)(x-t)^2+x^2)}dx\\
&= \lim_{t\to\infty} \frac{k^2}{k!(2\pi)^{k/2}} \int_{\mathbb{R}} \left(\frac{t(1+O((t-x)^{-2}))}{t/k-(x-(1-1/k)t)}\right)^{k-1} e^{-\frac{k}{2}(x-(1-1/k)t)^2}dx\\
&= \frac{k^{k+1/2}}{k!(2\pi)^{(k-1)/2}}.
\end{aligned}$$

(b) For fixed $\varepsilon > 0$ write $\mathbb{P}_\varepsilon$ for the conditional distribution of $\varepsilon^{-1}(\bar{Z}-Z_1,\bar{Z}-Z_2,\ldots,\bar{Z}-Z_k)$ given $\mathcal{B}_\varepsilon := \{\bar{Z}-Z_{\min}\le\varepsilon\}$. For any $\varepsilon>0$, the support of $\mathbb{P}_\varepsilon$ is the set $\Delta_k := \{(x_1,x_2,\ldots,x_k) \mid x_1+x_2+\cdots+x_k=0, x_1\le 1,\ldots,x_k\le 1\}$ which is a simplex with corner points $v_i := (1,\ldots,1,1-k,1,\ldots,1)$ where $1-k$ is in the i-th position for $i=1,2,\ldots,k$. The distribution is coordinate permutation invariant for each $\varepsilon>0$. Obviously $\{\mathbb{P}_\varepsilon : \varepsilon>0\}$ is a tight family of probability measures on $\mathbb{R}^k$. Every subsequential limit of $\mathbb{P}_\varepsilon$ as $\varepsilon\downarrow 0$ is translation invariant within the simplex. Hence the limiting distribution exists and is uniform over the simplex Δ_k. Now given a Dirichlet$(1,1,\ldots,1)$ random vector $\mathbf{U}=(U_1,U_2,\ldots,U_k)$ one can get a uniform random point from the simplex Δ_k by taking $\sum_{i=1}^k U_i v_i = (1-kU_1, 1-kU_2,\ldots,1-kU_k)$. Thus we are done.

(c) From the fact that $\bar{Z}-Z_{\min} = \max_{1\le i\le k}\{\bar{Z}-Z_i\}$ it is easy to see that

$$\mathbb{P}(\bar{Z}-Z_1\ge x)\le \mathbb{P}(\bar{Z}-Z_{\min}\ge x)\le k\,\mathbb{P}(\bar{Z}-Z_1\ge x).$$

The rest follows from the fact $\bar{Z}-Z_1$ is normal with mean zero and variance $(k-1)/k$, and for a standard Gaussian random variable Z,

$$\mathbb{P}(Z\ge x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}x}(1+o(1)) \text{ as } x\uparrow\infty. \qquad \blacksquare$$

5.2. **Maxima of two correlated gaussian r.v.s.** Let (Z, Z_ρ) be a bivariate gaussian random vector with $\mathbb{E}(Z)=\mathbb{E}(Z_\rho)=0, \operatorname{Var}(Z)=\operatorname{Var}(Z_\rho)=1$ and $\mathbb{E}(ZZ_\rho)=\rho\ge 0$. Note that, we can explicitly construct such a distribution by taking Z be standard normal and constructing $Z_\rho = \rho Z+\sqrt{1-\rho^2}Z'$ where Z' is an i.i.d. copy of Z. To estimate the conditional probability $\mathbb{P}(Z_\rho > x \mid Z > x)$ for 'large' x, first of all note that conditional on $\mathcal{A}=\{Z>x\}$, the random variable $x(Z-x)$ remains tight (in fact, using an argument similar to the proof of Lemma 5.2(b), one can show it converges to Exp(1) distribution as $x\to\infty$) and thus Z is concentrated around x conditional on $\mathcal{A}$. Now conditionally on $\mathcal{A}$, the event $\{\rho Z+\sqrt{1-\rho^2}Z'>x\}$ is roughly equivalent to the event $\{\sqrt{1-\rho^2}Z' > (1-\rho)x\}$, which has probability $\bar{\Phi}(\theta x)$ where $\theta=\sqrt{(1-\rho)/(1+\rho)}$. The following Lemma 5.4 makes these ideas precise.

Lemma 5.4 (eqn. (1.2) in [36]). *Let Z, Z' be two i.i.d. standard Gaussian r.v.s. Then, for any $\rho\in[0,1]$ and $x>0$ we have*

$$\bar{\Phi}(\theta x)\leqslant \mathbb{P}(\rho Z+\sqrt{1-\rho^2}Z'>x \mid Z>x)\leqslant (1+\rho)\bar{\Phi}(\theta x)$$

where

$$\theta=\sqrt{\frac{1-\rho}{1+\rho}}.$$

5.3. **Proof of Lemma 3.10.** Recall that Lemma 3.10 computed the asymptotic probability of the maximum row and column sum of a rectangular matrix being large simultaneously. Using the tail bounds in the previous Section, let us now prove this Lemma.

Proof. Let $Z, Z_1, Z_2, \ldots, Z_s, Z_1', Z_2', \ldots, Z_t'$ be i.i.d. standard Gaussian r.v.s. Define

$$V_s = \max_{1\le i\le s}(Z_i - \bar{Z}), \qquad V_t' = \max_{1\le j\le t}(Z_i' - \bar{Z}').$$

It is easy to see that

$$(\max_{1\le i\le s} u_{i\cdot} - u_{\cdot\cdot}, \max_{1\le j\le t} u_{\cdot j} - u_{\cdot\cdot}, u_{\cdot\cdot}) \overset{\mathrm{d}}{=} (t^{-1/2}V_s, s^{-1/2}V_t', (st)^{-1/2}Z).$$

Define $\alpha_n = \sqrt{st}(\theta b_n + x/a_n)$ and $\beta_n = \sqrt{st}(\theta b_n + y/a_n)$. Hence we have

$$\begin{aligned} p :=& \mathbb{P}(\max_{1\le i\le s} u_{i\cdot} \ge \theta b_n + x/a_n, \max_{1\le j\le t} u_{\cdot j} \ge \theta b_n + y/a_n) \\ =& \mathbb{P}(V_s \ge (\alpha_n - Z)/\sqrt{s}, V_t' \ge (\beta_n - Z)/\sqrt{t}). \end{aligned}$$

Now using Lemma 5.3(c) we have

$$p = (\sqrt{st}g_s g_t + o(1))\,\mathbb{E}\left((\alpha_n - Z)^{-1}(\beta_n - Z)^{-1}\exp\left(-\frac{(\alpha_n - Z)^2}{2(s-1)} - \frac{(\beta_n - Z)^2}{2(t-1)}\right)\right)$$

where the expectation is w.r.t. Z. One can easily check that

$$\begin{aligned} &\frac{(\alpha_n - z)^2}{s-1} + \frac{(\beta_n - z)^2}{t-1} + z^2 \\ &\quad = \frac{st-1}{(s-1)(t-1)}\left(z - \frac{(t-1)\alpha_n + (s-1)\beta_n}{st-1}\right)^2 + \frac{t\alpha_n^2 + s\beta_n^2 - 2\alpha_n\beta_n}{st-1}. \end{aligned}$$

Define

$$q := \exp\left(-\frac{t\alpha_n^2 + s\beta_n^2 - 2\alpha_n\beta_n}{2(st-1)}\right).$$

Recalling that $a_n, b_n \sim \sqrt{2\log n}$, we have

$$\begin{aligned} p &= \frac{q(st-1)^2(\sqrt{st}g_s g_t + o(1))}{(s-1)(t-1)(t\alpha_n - \beta_n)(s\beta_n - \alpha_n)}\int_{\mathbb{R}} \exp\left(-\frac{(st-1)z^2}{2(s-1)(t-1)}\right)dz \\ &= \frac{(\eta(s,t,\theta) + o(1))}{\log n}q \end{aligned}$$

for some constant $\eta(s,t,\theta) > 0$. Finally note that,

$$q = \exp\left(-\frac{st\theta((t-1)x + (s-1)y)}{st-1}\right)\left(\frac{\sqrt{4\pi\log n}}{n}\right)^{\frac{st(s+t-2)\theta^2}{st-1}}(1 + o(1))$$

and this completes the proof. ■

5.4. **Combinatorial estimates.** We start with the following Lemma on binomial coefficients which easily follows from Stirling's approximation.

Lemma 5.5. *For any $n, 1 \le k \le \sqrt{n}$ we have*

$$\frac{(n)_k}{n^k} = e^{-k^2/2n + O(k/n)}.$$

The next shows the asymptotic negligibility of a particular series which arises in deriving results about the global optima.

Lemma 5.6. (a) *Let $N = \binom{n}{k}^2, a_N = \sqrt{2\log N}, b_N = \sqrt{2\log N} - \log(4\pi \log N)/2\sqrt{2\log N}$ and $u_N = b_N - x/a_N$. There exists a universal constant $c > 0$, such that for $k \le c\log n/\log\log n$ and any fixed $x \in \mathbb{R}$, we have*

$$\sum_{\substack{1\le s,t\le k\\ st\neq k^2}} \binom{k}{s}\binom{k}{t}\binom{n-k}{k-s}\binom{n-k}{k-t}\binom{n}{k}^{-2}\sqrt{\frac{k^2+st}{k^2-st}} \cdot e^{stu_N^2/(k^2+st)} \to 0$$

as $n \to \infty$.

(b) *The same result holds if $\log k \ll \log n$ and $k(\log\log n)^2/\log n \ll x \ll a_N^2$.*

Proof of Lemma 5.6. Throughout the analysis we will assume that $k = e^{o(\log n)}$. Let $N = \binom{n}{k}^2$ and define $a_N = \sqrt{2\log N}, b_N = \sqrt{2\log N} - \log(4\pi\log N)/2\sqrt{2\log N}$.

Let $u_N = b_N - x_n/a_N$ with $-K \le x_n \ll a_N^2$ for some constant $K \ge 1$. Clearly we have $e^{u_N^2/2} = Ne^{-x_n r_n + o(1)}/\sqrt{2\pi}a_N$ where $r_n = 1 + x_n/2a_N^2 \to 1$ as $n \to \infty$. Moreover using Stirling's approximation and the fact that $k \ll \sqrt{n}$ one can easily see that $N = \frac{1+o(1)}{2\pi k}(en/k)^{2k}$. Thus we have

$$e^{u_N^2/4k} \le \frac{en}{k}\left(\frac{c_0 e^{-x_n r_n}}{\sqrt{k^3\log n}}\right)^{1/2k} \tag{5.2}$$

for some universal constant $c_0 > 0$. Now note that $(k^2+st)/(k^2-st) \le k$ for all integers s, t with $2 \le s+t \le 2k-1$. Thus we need to show that

$$I_n := \sqrt{k}\sum_{1\le s,t\le k, st\neq k^2}\binom{k}{s}\binom{k}{t}\binom{n-k}{k-s}\binom{n-k}{k-t}\binom{n}{k}^{-2} e^{stu_N^2/(k^2+st)} \to 0$$

as $n \to \infty$.

Note that

$$\binom{n-k}{k-s}\binom{n}{k}^{-1} = \frac{(k)_s(n)_{2k-s}}{(n)_k^2}$$

where $(n)_k := n!/(n-k)!$. Thus $k \ll \sqrt{n}$ and Lemma 5.5 imply that

$$\binom{n-k}{k-s}\binom{n-k}{k-t}\binom{n}{k}^{-2} \le \frac{c(k)_s(k)_t}{n^{s+t}}$$

for some universal constant $c > 0$. Using Stirling's formula once more, we have

$$(k)_s(k)_t \le k(k/e)^{s+t}e^{-k(1-s/t)\log(1-s/k) - k(1-t/k)\log(1-t/k))}.$$

Moreover the function $f(x) = -(1-x)\log(1-x)$ is concave. Thus we have

$$(k)_s(k)_t \le k(k/e)^{s+t}e^{2kf((s+t)/2k)}.$$

Note that $4st \le (s+t)^2$ implies that

$$\frac{st}{k^2+st} \le \frac{(s+t)^2}{4k^2+(s+t)^2}.$$

Thus

$$I_n \le ck^{3/2} \sum_{\substack{1\le s,t\le k\\ st\neq k^2}} \binom{k}{s}\binom{k}{t}\left(\frac{k}{en}\right)^{s+t} \exp\left(2kf\left(\frac{s+t}{2k}\right) + \frac{u_N^2(s+t)^2}{4k^2+(s+t)^2}\right).$$

Now

$$\sum_{s=0}^{l}\binom{k}{s}\binom{k}{l-s} = \binom{2k}{l} \le \exp(2kh(l/2k))$$

for all integer l where $h(x) = -x\log x - (1-x)\log(1-x)$. Thus we have

$$\begin{aligned} I_n &\le ck^{3/2}\sum_{l=2}^{2k-1}\left(\frac{k}{en}\right)^{l}\exp\left(2kg\left(\frac{l}{2k}\right) + \frac{u_N^2 l^2}{4k^2+l^2}\right)\\ &= ck^{3/2}\sum_{l=2}^{2k-1}\left(\frac{ke^{u_N^2/4k}}{en}\right)^{l}\exp\left(2kg\left(\frac{l}{2k}\right) - \frac{u_N^2 l(2k-l)^2}{4k(4k^2+l^2)}\right) \end{aligned}$$

where $g(x) = -x\log x - 2(1-x)\log(1-x)$. Using Equation (5.2) we have

$$\frac{ke^{u_N^2/4k}}{en} \le \left(\frac{c_0e^{-x_nr_n}}{\sqrt{k^3\log n}}\right)^{1/2k}$$

and thus finally we have

$$\begin{aligned} I_n &\le c_1k^{3/2}\sum_{l=2}^{2k-1}\exp\left(2kg\left(\frac{l}{2k}\right) - \frac{u_N^2 l(2k-l)^2}{4k(4k^2+l^2)}\right)\left(\frac{e^{-x_nr_n}}{\sqrt{k^3\log n}}\right)^{l/2k}\\ &\le c_1k^{3/2}\sum_{l=2}^{2k-1}\exp\left(2kg\left(\frac{l}{2k}\right) - \frac{u_N^2}{2(1+(l/2k)^2)}\frac{l}{2k}\left(1-\frac{l}{2k}\right)^2\right)\left(\frac{e^{-x_nr_n}}{\sqrt{k^3\log n}}\right)^{l/2k}. \end{aligned}$$

Consider the function

$$\psi(x) := \frac{2g(x)}{x(1-x)^2(1+x^2)^{-1}}, x\in(0,1)$$

which is positive, convex and diverges to infinity as $x \to 0$ or 1. Moreover for $x \in [1/k, 3/4]$ we have $\psi(x) \le c\log k$ for some constant $c > 0$. Now, it is easy to see that $c\log k \le \frac{u_N^2}{4k}$ for $k = e^{o(\log n)}$. Thus we have

$$\begin{aligned} &k^{3/2}\sum_{l=2}^{3k/2}\exp\left(2kg\left(\frac{l}{2k}\right) - \frac{u_N^2}{2(1+(l/2k)^2)}\frac{l}{2k}\left(1-\frac{l}{2k}\right)^2\right)\left(\frac{e^{-x_nr_n}}{\sqrt{k^3\log n}}\right)^{l/2k}\\ &\le k^{3/2}\sum_{l=2}^{3k/2}\exp\left(-\frac{u_N^2}{2^8k}\cdot l\right) \le \frac{k^{3/2}\exp(-u_N^2/2^7k)}{1-\exp(-u_N^2/2^8k)} \to 0 \end{aligned}$$

as $n \to \infty$ as $u_N^2/k \approx \log n$.

Now the function $\psi(x) \leq c' k \log k$ for $x \in [3/4, 1 - 1/2k]$ and $c' k \log k \leq u_N^2/4k$ only when $k \leq c \log n / \log\log n$ for some small constant $c > 0$. In that case, we have

$$
\begin{aligned}
&k^{3/2} \sum_{l=3k/2}^{2k-1} \exp\left(2kg\left(\frac{l}{2k}\right) - \frac{u_N^2}{2(1+(l/2k)^2)} \frac{l}{2k}\left(1 - \frac{l}{2k}\right)^2\right)\left(\frac{e^{-x_n r_n}}{\sqrt{k^3 \log n}}\right)^{l/2k} \\
&\leq k^{3/2} \sum_{3k/2}^{2k-1} \exp\left(-\frac{u_N^2}{2^6 k^2}(2k-l)^2\right) \frac{1}{(k^3 \log n)^{3/8}} \\
&\leq \frac{k^{3/8}}{(\log n)^{3/8}} \sum_{l=1}^{k/2} \exp\left(-\frac{u_N^2}{2^6 k^2} l^2\right) \to 0
\end{aligned}
$$

as $\log n/k, u_N^2/k^2 \to \infty$ as $n \to \infty$.

Now when $k \geq c \log n / \log\log n$, we need to break the second sum into into two parts and take $x_n \gg k(\log\log n)^2/\log n$. For $x \in [3/4, 1 - \log\log n/(2c \log n)]$ we have $\psi(x) \leq c' \log n \leq u_N^2/4k$ and thus

$$
\begin{aligned}
&k^{3/2} \sum_{l=3k/2}^{2k(1-\log\log n/(2c\log n))} \exp\left(2kg\left(\frac{l}{2k}\right) - \frac{u_N^2}{2(1+(l/2k)^2)} \frac{l}{2k}\left(1 - \frac{l}{2k}\right)^2\right)\left(\frac{e^{-x_n r_n}}{\sqrt{k^3 \log n}}\right)^{l/2k} \\
&\leq k^{3/2} \sum_{3k/2}^{2k(1-\log\log n/(2c\log n))} \exp\left(-\frac{u_N^2}{2^6 k^2}(2k-l)^2\right) \frac{e^{-3x_n/4}}{(k^3 \log n)^{3/8}} \\
&\leq \frac{k^{3/8} e^{-3x_n/4}}{(\log n)^{3/8}} \sum_{l=k\log\log n/c\log n}^{k/2} \exp\left(-\frac{u_N^2}{2^6 k^2} l^2\right) \\
&\leq \frac{k^{1+3/8} e^{-3x_n/4}}{(\log n)^{3/8}} \exp\left(-\frac{c' k (\log\log n)^2}{\log n}\right) \to 0
\end{aligned}
$$

as $n \to \infty$. For the last part we have

$$
\begin{aligned}
&k^{3/2} \sum_{l=2k(1-\log\log n/(2c\log n))}^{2k-1} \exp\left(2kg\left(\frac{l}{2k}\right) - \frac{u_N^2}{2(1+(l/2k)^2)} \frac{l}{2k}\left(1 - \frac{l}{2k}\right)^2\right)\left(\frac{e^{-x_n r_n}}{\sqrt{k^3 \log n}}\right)^{l/2k} \\
&\leq k^{5/2} \exp\left(2kg\left(1 - \log\log n/2c\log n\right)\right) e^{-x_n(1+o(1))} (k^3 \log n)^{-1/2+o(1)} \\
&\leq k^{5/2} \exp\left(c' k (\log\log n)^2/\log n - x_n(1+o(1))\right) (k^3 \log n)^{-1/2+o(1)} \to 0
\end{aligned}
$$

as $n \to \infty$. Thus we are done. ■

6. Proofs: Global optima

6.1. **Distributional results.** We start by proving the comparison result, Lemma 3.2 using which we shall complete the proof Theorem 3.7.

Proof of Lemma 3.2. Let Σ_1 be the matrix $((\sigma_{ij}))_{i,j=1}^N$ where $\sigma_{ii} = 1, i = 1, 2, \ldots, N$ and Σ_0 be the $N \times N$ identity matrix. For $t = 0, 1$ construct independent random variables $\mathbf{X}^t \sim \mathrm{N}(\mathbf{0}, \Sigma_t)$. For a general real number $t \in [0, 1]$ let

$$\mathbf{X}^t := \sqrt{t}\mathbf{X}^1 + \sqrt{1-t}\mathbf{X}^0 \tag{6.1}$$

Note that

$$\mathbf{X}^t = (X_1^t, X_2^t, \ldots, X_N^t) \sim \mathrm{N}(\mathbf{0}, \Sigma_t)$$

where $\Sigma_t = t\Sigma_1 + (1-t)\Sigma_0$.

Fix a smooth function $G(\mathbf{x})$ of N variables $\mathbf{x} = (x_1, x_2, \ldots, x_N)$. Define $G_{ij}(\mathbf{x}) = \frac{\partial^2 G}{\partial x_i \partial x_j}(\mathbf{x})$ for $1 \leq i, j \leq N$. Using the representation (6.1) and integration by parts one can easily prove that

$$\mathbb{E}(G(\mathbf{X}^1)) - \mathbb{E}(G(\mathbf{X}^0)) = \sum_{i<j} \sigma_{ij} \int_0^1 \mathbb{E}(G_{ij}(\mathbf{X}^t))dt.$$

We briefly sketch the proof for completeness. Note that $\mathbf{X}^t \overset{\mathrm{d}}{=} \Sigma_t^{1/2}\mathbf{X}^0$. Thus

$$\begin{aligned}
\mathbb{E}(G(\mathbf{X}^1)) - \mathbb{E}(G(\mathbf{X}^0)) &= \int_0^1 \frac{d}{dt}\, \mathbb{E}(G(\Sigma_t^{1/2}\mathbf{X}^0))dt \\
&= \int_0^1 \mathbb{E}(\nabla G(\Sigma_t^{1/2}\mathbf{X}^0)\frac{d}{dt}(\Sigma_t^{1/2})\mathbf{X}^0)dt \\
&= \int_0^1 \mathbb{E}(\mathrm{tr}(\frac{d}{dt}(\Sigma_t^{1/2})\nabla^2 G(\Sigma_t^{1/2}\mathbf{X}^0)\Sigma_t^{1/2})dt \\
&= \frac{1}{2}\int_0^1 \mathbb{E}(\mathrm{tr}(\frac{d}{dt}(\Sigma_t)\nabla^2 G(\mathbf{X}^t))dt \\
&= \frac{1}{2}\int_0^1 \mathbb{E}(\mathrm{tr}((\Sigma_1 - \Sigma_0)\nabla^2 G(\mathbf{X}^t))dt = \sum_{i<j}\sigma_{ij}\int_0^1 \mathbb{E}(G_{ij}(\mathbf{X}^t))dt.
\end{aligned}$$

Fix $\varepsilon > 0$. Take $G^\varepsilon(\mathbf{x}) = \prod_{i=1}^N \Phi(\varepsilon^{-1}(u - x_i))$ in place of G, where Φ is the normal distribution function and $\phi = \Phi'$. We have

$$\begin{aligned}
&|\,\mathbb{E}(G^\varepsilon(\mathbf{X}^1)) - \mathbb{E}(G^\varepsilon(\mathbf{X}^0))| \\
&\qquad \leq \sum_{i<j}|\sigma_{ij}|\int_0^1 \mathbb{E}(\varepsilon^{-2}\phi(\varepsilon^{-1}(u - X_i^t))\phi(\varepsilon^{-1}(u - X_j^T)))dt \\
&\qquad = \sum_{i<j}|\sigma_{ij}|\int_0^1 \mathbb{E}(\rho_{ij}^t(u + \varepsilon Z_1, u + \varepsilon Z_2))dt
\end{aligned}$$

where $\rho_{ij}^t(x,y)$ is the joint density of (X_i^t, X_j^t) and Z_1, Z_2 are i.i.d. Gaussian r.v.s. Taking $\varepsilon \to 0$ and noting that

$$G^\varepsilon(\mathbf{x}) \to \mathbb{1}\left\{\max_{1\leqslant i\leqslant N} x_i \leqslant u\right\},$$

we have

$$|\,\mathbb{P}(\max_{1\leq i\leq N} X_i \leq u) - \mathbb{P}(\max_{1\leq i\leq N} Z_i \leq u)| \leq \sum_{i<j}|\sigma_{ij}|\int_0^1 \rho_{ij}^t(u,u)dt.$$

Now

$$\rho_{ij}^t(u,u) = \frac{e^{-u^2/(1+\sigma_{ij}t)}}{2\pi\sqrt{1 - \sigma_{ij}^2 t^2}}$$

for $u \in \mathbb{R}$ and thus

$$|\sigma_{ij}| \int_0^1 \rho_{ij}^t(u,u)dt \le \left| \int_0^{\sigma_{ij}} \frac{e^{-u^2/(1+t)}}{2\pi\sqrt{1-t^2}} dt \right|.$$

Changing the variable t to $x = \sqrt{(1-t)/(1+t)}$ we have

$$\begin{aligned} |\sigma_{ij}| \int_0^1 \rho_{ij}^t(u,u)dt &\le \pi^{-1} e^{-u^2/2} \left| \int_{\theta_{ij}}^1 \frac{e^{-(ux)^2/2}}{1+x^2} dx \right| \\ &\le \pi^{-1} e^{-u^2/2} \left| \int_{\theta_{ij}}^1 e^{-(ux)^2/2} dx \right| \\ &\le 2\bar{\Phi}(u)|\bar{\Phi}(u) - \bar{\Phi}(\theta_{ij}u)| \\ &\le 2\bar{\Phi}(u)\bar{\Phi}(\theta_{ij}^1 u)\min\{1, |1-\theta_{ij}|u\phi(\theta_{ij}^1 u)/\bar{\Phi}(\theta_{ij}^1 u)\} \\ &\le 2\bar{\Phi}(u)\bar{\Phi}(\theta_{ij}^1 u)\min\{1, |1-\theta_{ij}|u(u+1)\} \end{aligned}$$

where $\theta_{ij} = \sqrt{(1-\sigma_{ij})/(1+\sigma_{ij})}$, $\theta_{ij}^1 = \min\{\theta_{ij}, 1\}$ and we used the fact that $\Phi(x)$ is concave and $\bar{\Phi}(x) \ge \phi(x)/(1+x)$ for $x \ge 0$. This completes the proof. ■

Let us now prove asymptotics for the global maximum.

Proof of Theorem 3.1: We start with the proof of part (a). Recall that $|\mathscr{S}_n(k)| = N = \binom{n}{k}^2$. For any $x \in \mathbb{R}$, we have

$$\mathbb{P}(a_N(k \operatorname{avg}(\mathbf{W}_{\lambda^*}) - b_N) \le x) = \mathbb{P}(\max_{\lambda \in \mathscr{S}_n(k)} k \operatorname{avg}(\mathbf{W}_\lambda) \le b_N + a_N^{-1}x).$$

Moreover, $k \operatorname{avg}(\mathbf{W}_\lambda) \sim \mathrm{N}(0,1)$ for all $\lambda \in \mathscr{S}_n(k)$. Further for i.i.d. standard Gaussian r.v.s $Z_1, Z_2, \ldots, Z_N$, standard results in extreme value theory (see Lemma 3.6) imply that for any $x \in \mathbb{R}$,

$$\mathbb{P}(\max_{1\le i\le N} Z_i \le b_N + a_N^{-1}x) \to e^{-e^{-x}} = \mathbb{P}(-\log T \le x) \text{ as } N \to \infty$$

where $T \sim \operatorname{Exp}(1)$. Thus it is enough to show that

$$\left| \mathbb{P}(\max_{\lambda \in \mathscr{S}_n(k)} k \operatorname{avg}(\mathbf{W}_\lambda) \le u_N) - \mathbb{P}(\max_{1\le i\le N} Z_i \le u_N) \right| \to 0$$

as $N \to \infty$ where $u_N = b_N + a_N^{-1}x$. First note that $\operatorname{Cov}(k \operatorname{avg}(\mathbf{W}_\lambda), k \operatorname{avg}(\mathbf{W}_{\lambda'})) = stk^{-2}$ where $|\lambda \cap \lambda'| = (s,t)$, *i.e.*, λ, λ' share s many rows and t many columns.

Let $u_N = b_N + a_N^{-1}x$. Using Proposition 3.2 we need to show that

$$N^2 \bar{\Phi}(u_N)^2 \sum_{\substack{1\le s,t\le k \\ st \ne k^2}} \binom{k}{s}\binom{k}{t}\binom{n-k}{k-s}\binom{n-k}{k-t}\binom{n}{k}^{-2} \sqrt{\frac{k^2+st}{k^2-st}} \cdot e^{stu_N^2/(k^2+st)} \to 0$$

It is easy to check that $N\bar{\Phi}(u_N) \to e^{-x}$ as $N \to \infty$. Lemma 5.6(a) completes the proof.

Part (b) follows in an identical fashion using Slepian's lemma and Lemma 5.6(b). We omit the details.

Now we move to the proof of part (c). For k fixed, from part (a) we have for any fixed $x \in \mathbb{R}$,

$$\mathbb{P}(a_N(kM_n(k) - b_N) \ge x) \to \mathbb{P}(-\log T \ge x)$$

as $n \to \infty$ where T is an exponential rate one random variable and $N = \binom{n}{k}^2 = |\mathscr{S}_n(k)|$. Fix a subset $A \subseteq \mathbb{R}^{k\times k}$ and $x \in \mathbb{R}$. Define

$$x_n = (b_N + x/a_N)/k.$$

Given a matrix $\mathbf{U}$, we will write $\hat{\mathbf{U}}$ to denote the matrix $\mathbf{U} - \operatorname{avg}(\mathbf{U})\mathbf{1}\mathbf{1}'$. It is enough to show that for any Borel set $A \in \mathcal{B}(\mathbb{R}^{k\times k})$,

$$\begin{aligned}\mathbb{P}(M_n(k) \geq x_n, &\mathbf{W}_{\lambda^*(k)} - \operatorname{avg}(\mathbf{W}_{\lambda^*(k)})\mathbf{1}\mathbf{1}' \in A) \\ &- \mathbb{P}(M_n(k) \geq x_n)\,\mathbb{P}(\hat{\mathbf{W}}_{[k]\times[k]} \in A) \to 0\end{aligned}$$

as $n \to \infty$. We have by symmetry

$$\begin{aligned}p :=& \,\mathbb{P}(M_n(k) \geq x_n, \mathbf{W}_{\lambda^*(k)} - \operatorname{avg}(\mathbf{W}_{\lambda^*(k)})\mathbf{1}\mathbf{1}' \in A) \\ =& \sum_{\gamma\in\mathscr{S}_n(k)} \mathbb{P}(\lambda^*(k) = \gamma, \operatorname{avg}(\mathbf{W}_\gamma) \geq x_n, \hat{\mathbf{W}}_\gamma \in A) \\ =& N\,\mathbb{P}(\lambda^*(k) = \lambda, \operatorname{avg}(\mathbf{W}_\lambda) \geq x_n, \hat{\mathbf{W}}_\lambda \in A)\end{aligned}$$

where $\lambda = [k] \times [k]$ is the corner submatrix. Define $\mathscr{E}_\lambda := \{\gamma \in \mathscr{S}_n(k) \mid \gamma \neq \lambda, \lambda \cap \gamma = \emptyset\}$, namely the set of submatrices which have no overlap with λ and $\mathscr{N}_\lambda = \{\gamma \in \mathscr{S}_n(k) \mid \gamma \neq \lambda, \lambda \cap \gamma \neq \emptyset\}$ so that the entire configuration space $\mathscr{S}_n(k) = \mathscr{E}_\lambda \cup \mathscr{N}_\lambda$. Now the event

$$\{\lambda^*(k) = \lambda\} = \{\max_{\gamma\in\mathscr{E}_\lambda} \operatorname{avg}(\mathbf{W}_\gamma) < \operatorname{avg}(\mathbf{W}_\lambda), \max_{\gamma\in\mathscr{N}_\lambda} \operatorname{avg}(\mathbf{W}_\gamma) < \operatorname{avg}(\mathbf{W}_\lambda)\}.$$

Moreover, $\max_{\gamma\in\mathscr{E}_\lambda} \operatorname{avg}(\mathbf{W}_\gamma), \operatorname{avg}(\mathbf{W}_\lambda)$ and $\hat{\mathbf{W}}_\lambda$ are independent. Thus

$$\begin{aligned}&N^{-1}p \\ &= \mathbb{P}(\max_{\gamma\in\mathscr{E}_\lambda} \operatorname{avg}(\mathbf{W}_\gamma) < \operatorname{avg}(\mathbf{W}_\lambda), \max_{\gamma\in\mathscr{N}_\lambda} \operatorname{avg}(\mathbf{W}_\gamma) < \operatorname{avg}(\mathbf{W}_\lambda), \operatorname{avg}(\mathbf{W}_\lambda) \geq x_n, \hat{\mathbf{W}}_\lambda \in A) \\ &= \mathbb{P}(\max_{\gamma\in\mathscr{E}_\lambda} \operatorname{avg}(\mathbf{W}_\gamma) < \operatorname{avg}(\mathbf{W}_\lambda), \operatorname{avg}(\mathbf{W}_\lambda) \geq x_n)\,\mathbb{P}(\hat{\mathbf{W}}_\lambda \in A) \\ &\ - \mathbb{P}(\max_{\gamma\in\mathscr{E}_\lambda} \operatorname{avg}(\mathbf{W}_\gamma) < \operatorname{avg}(\mathbf{W}_\lambda), \max_{\gamma\in\mathscr{N}_\lambda} \operatorname{avg}(\mathbf{W}_\gamma) \geq \operatorname{avg}(\mathbf{W}_\lambda), \operatorname{avg}(\mathbf{W}_\lambda) \geq x_n, \hat{\mathbf{W}}_\lambda \in A).\end{aligned}$$

In particular we have

$$\begin{aligned}&\left| N^{-1}p - \mathbb{P}(\hat{\mathbf{W}}_\lambda \in A)\,\mathbb{P}(\max_{\gamma\in\mathscr{E}_\lambda} \operatorname{avg}(\mathbf{W}_\gamma) < \operatorname{avg}(\mathbf{W}_\lambda), \operatorname{avg}(\mathbf{W}_\lambda) \geq x_n) \right| \\ &\quad \leq \mathbb{P}(\max_{\gamma\in\mathscr{N}_\lambda} \operatorname{avg}(\mathbf{W}_\gamma) \geq \operatorname{avg}(\mathbf{W}_\lambda) \geq x_n).\end{aligned}$$

Since the upper bound does not depend on the set A, taking $A = \mathbb{R}^{k\times k}$ and simplifying we have

$$\begin{aligned}&\left| \mathbb{P}(M_n(k) \geq x_n, \mathbf{W}_{\lambda^*(k)} - \operatorname{avg}(\mathbf{W}_{\lambda^*(k)})\mathbf{1}\mathbf{1}' \in A) - \mathbb{P}(M_n(k) \geq x_n)\,\mathbb{P}(\hat{\mathbf{W}}_\lambda \in A) \right| \\ &\leq 2N\,\mathbb{P}(\max_{\gamma\in\mathscr{N}_\lambda} \operatorname{avg}(\mathbf{W}_\gamma) \geq \operatorname{avg}(\mathbf{W}_\lambda) \geq x_n) \\ &= 2N\,\mathbb{P}(\operatorname{avg}(\mathbf{W}_\lambda) \geq x_n) \cdot \mathbb{P}(\max_{\gamma\in\mathscr{N}_\lambda} \operatorname{avg}(\mathbf{W}_\gamma) \geq \operatorname{avg}(\mathbf{W}_\lambda) \mid \operatorname{avg}(\mathbf{W}_\lambda) \geq x_n).\end{aligned}$$

Note that $N\,\mathbb{P}(\operatorname{avg}(\mathbf{W}_\lambda) \geq x_n) \to e^{-x}$ as $n \to \infty$. Thus we have to show that

$$\mathbb{P}(\max_{\gamma\in\mathscr{N}_\lambda} \operatorname{avg}(\mathbf{W}_\gamma) \geq \operatorname{avg}(\mathbf{W}_\lambda) \mid \operatorname{avg}(\mathbf{W}_\lambda) \geq x_n) \to 0$$

as $n \to \infty$. Let $\mathscr{N}_\lambda(s,t) = \{\gamma \in \mathscr{S}_n(k) \mid |\lambda \cap \gamma| = (s,t)\}, 1 \le s,t \le k$. Clearly $|\mathscr{N}_\lambda(s,t)| = \binom{k}{s}\binom{k}{t}\binom{n-k}{k-s}\binom{n-k}{k-t} = O(n^{2k-s-t})$. The union bound gives

$$\begin{aligned}
&\mathbb{P}(\max_{\gamma\in\mathscr{N}_\lambda} \operatorname{avg}(\mathbf{W}_\gamma) \ge \operatorname{avg}(\mathbf{W}_\lambda) \mid \operatorname{avg}(\mathbf{W}_\lambda) \ge x_n) \\
&\le \sum_{1\le s,t\le k, st\neq k^2} \mathbb{P}(\max_{\gamma\in\mathscr{N}_\lambda(s,t)} \operatorname{avg}(\mathbf{W}_\gamma) \ge \operatorname{avg}(\mathbf{W}_\lambda) \mid \operatorname{avg}(\mathbf{W}_\lambda) \ge x_n).
\end{aligned}$$

Note that $\hat{\mathbf{W}}_\lambda := \mathbf{W}_\lambda - \operatorname{avg}(\mathbf{W}_\lambda)\mathbf{1}\mathbf{1}' = ((\hat{w}_{ij}))$ is independent of $\operatorname{avg}(\mathbf{W}_\lambda)$ and we have $\max_{i,j\in[k]} w_{ij} = \operatorname{avg}(\mathbf{W}_\lambda) + \max_{i,j\in[k]} \hat{w}_{ij}$.

For each $\gamma \in \mathscr{N}_\lambda(s,t)$, $\mathbf{W}_\gamma$ has exactly $k^2 - st$ many elements outside the submatrix $\mathbf{W}_\lambda$ and these entries are independent of the matrix $\mathbf{W}_\lambda$. Let $F(\mathbf{W}_\gamma)$ be the average of the ($k^2 - st$ many) entries in $\mathbf{W}_\gamma$ that are outside $\mathbf{W}_\lambda$. Thus $\max_{\gamma\in\mathscr{N}_\lambda(s,t)} \operatorname{avg}(\mathbf{W}_\gamma) \ge \operatorname{avg}(\mathbf{W}_\lambda)$ implies that $\max_{\gamma\in\mathscr{N}_\lambda(s,t)} F(\mathbf{W}_\gamma) \ge \operatorname{avg}(\mathbf{W}_\lambda) - (k^2-st)^{-1}st \max_{i,j\in[k]} \hat{w}_{ij}$. Thus

$$\begin{aligned}
&\mathbb{P}(\max_{\gamma\in\mathscr{N}_\lambda(s,t)} \operatorname{avg}(\mathbf{W}_\gamma) \ge \operatorname{avg}(\mathbf{W}_\lambda) \mid \operatorname{avg}(\mathbf{W}_\lambda) \ge x_n) \\
&\le \mathbb{P}(\max_{\gamma\in\mathscr{N}_\lambda(s,t)} F(\mathbf{W}_\gamma) \ge x_n - (k^2-st)^{-1}st \max_{i,j\in[k]} \hat{w}_{ij})
\end{aligned}$$

Using Slepian's lemma we have

$$\mathbb{P}(\max_{\gamma\in\mathscr{N}_\lambda(s,t)} F(\mathbf{W}_\gamma) \ge x) \le \mathbb{P}(V_{|\mathscr{N}_\lambda(s,t)|} \ge x\sqrt{k^2-st})$$

for all $x \in \mathbb{R}$ where we use V_n to denote the maximum of n many i.i.d. standard Gaussians. Thus

$$\begin{aligned}
&\mathbb{P}(\max_{\gamma\in\mathscr{N}_\lambda(s,t)} \operatorname{avg}(\mathbf{W}_\gamma) \ge \operatorname{avg}(\mathbf{W}_\lambda) \mid \operatorname{avg}(\mathbf{W}_\lambda) \ge x_n) \\
&\le \mathbb{P}(V_{|\mathscr{N}_\lambda(s,t)|} \ge x_n\sqrt{k^2-st} - (k^2-st)^{-1/2}st \max_{i,j\in[k]} \hat{w}_{ij}) \\
&\le |\mathscr{N}_\lambda(s,t)|\, \mathbb{P}(V_1 \ge x_n\sqrt{k^2-st} - (k^2-st)^{-1/2}st \max_{i,j\in[k]} \hat{w}_{ij}). \qquad (6.2)
\end{aligned}$$

Now, $|\mathscr{N}_\lambda(s,t)| = O(n^{2k-s-t})$ and $x_n\sqrt{k^2-st} \approx \sqrt{(2k-2st/k)2\log n}$. Now it is easy to see that $2st/k \le (s+t)^2/2k \le (1-1/2k)(s+t) \le s+t-1/k$. Thus the probability in (6.2) converges to zero as $n \to \infty$ and we are done. ∎

6.2. **Two point localization.** Here we shall prove Theorem 3.4. Fix $\tau > 0$ and recall the definition of $\tilde{k}$ from (3.1) as well as k^*. For any fixed m, let $N_n(m)$ denote the number of sub-matrices of size m with average greater than τ, *i.e.*,

$$N_n(m) := \sum_{\lambda\in\mathscr{S}_n(m)} \mathbb{1}\{\operatorname{avg}(\mathbf{W}_\lambda) > \tau\}.$$

Further note that if there is a sub-matrix of size m with average $> \tau$, then there exists a sub-matrix of size $m-1$ with average greater than τ. Thus the following Proposition 6.1 completes the proof. ∎

Proposition 6.1. *For fixed $\tau > 0$ we have the following asymptotics.*

(i) *Let $m = k^* + 1$. Then*

$$\mathbb{P}(N_n(m) > 0) \to 0 \text{ as } n \to \infty.$$

(ii) *Let $m = k^* - 1$. Then*

$$\mathbb{P}(N_n(m) > 0) \to 1 \text{ as } n \to \infty.$$

Proof of Proposition 6.1. Define

$$f_n(x) := \binom{n}{x}^2 \bar{\Phi}(x\tau)$$

for $x \in [1, n]$. It is easy to see that $\mathbb{E}(N_n(m)) = f_n(m)$. Moreover, using Lemma 5.1 and Stirling's approximation $\Gamma(x+1) = \sqrt{2\pi} x^{x+1/2} e^{-x+O(1/x)}$ for $x \geq 1$, for any constant $c \in \mathbb{R}$ we have (with $x = \tilde{k}$)

$$\begin{aligned}
&\frac{f_n(x+c)}{f_n(x)} \\
&= \left(\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(x+c+1)\Gamma(n-x-c+1)}\right)^2 \cdot \frac{\bar{\Phi}((x+c)\tau)}{\bar{\Phi}(x\tau)} \\
&= \frac{x^{2x+2}(n-x)^{2n-2x+1}}{(x+c)^{2x+2c+2}(n-x-c)^{2n-2x-2c+1}} \cdot e^{-(2cx+c^2)\tau^2/2+O(1/x)} \\
&= \frac{(1-x/n)^{2c}}{(1+c/x)^{2x+2c+2}(1-c/(n-x))^{2n-2x-2c+1}} \cdot \left(\frac{en}{x}e^{-x\tau^2}\right)^c e^{-c-c^2\tau^2/2+O(1/x)}.
\end{aligned}$$

Using (3.2) one can easily check that

$$\frac{en}{\tilde{k}} \exp(-\tilde{k}\tau^2/4) \to 1 \text{ as } n \to \infty.$$

Thus

$$\frac{f_n(\tilde{k}+c)}{f_n(\tilde{k})} \to 0 \text{ or } \infty \text{ as } n \to \infty$$

depending on whether $c > 0$ or $c < 0$. In particular, we have

$$\mathbb{E}(N_n(k^*+1)) \to 0 \text{ and } \mathbb{E}(N_n(k^*-1)) \to \infty \text{ as } n \to \infty.$$

Part (i) now follows easily from the fact that $\mathbb{P}(N_n(k^*+1) > 0) \leq \mathbb{E}(N_n(k^*+1))$. To prove (ii), we shall use the second moment method. To simplify notation, for the rest of this proof we write $k = k^* - 1$. We have already proved that $\mathbb{E}(N_n(k)) \to \infty$ as $n \to \infty$. By the second moment method, it is enough to show that

$$\frac{\mathbb{E}(N_n(k)^2)}{(\mathbb{E}\, N_n(k))^2} \to 1 \text{ as } n \to \infty.$$

Note that the collection of random variables $(\operatorname{avg}(\mathbf{W}_\lambda) : \lambda \in \mathscr{S}_n(k))$ is transitive, in the sense that for any $\lambda_0, \lambda_1 \in \mathscr{S}_n(k)$, there exists a permutation $\pi : \mathscr{S}_n(k) \to \mathscr{S}_n(k)$ with $\pi(\lambda_0) = \lambda_1$ such that

$$(\operatorname{avg}(\mathbf{W}_\lambda) : \lambda \in \mathscr{S}_n(k)) \stackrel{\mathrm{d}}{=} (\operatorname{avg}(\mathbf{W}_{\pi(\lambda)}) : \lambda \in \mathscr{S}_n(k)).$$

A simple calculation using this transitivity now implies that to prove the second assertion, it is enough to show for a fixed $\lambda_0 \in \mathscr{S}_n(k)$,

$$I_n := \sum_{\lambda \in \mathscr{S}_n(k), \lambda \neq \lambda_0, \lambda \cap \lambda_0 \neq \emptyset} \frac{\mathbb{P}(\operatorname{avg}(\mathbf{W}_\lambda) > \tau \mid \operatorname{avg}(\mathbf{W}_{\lambda_0}) > \tau)}{\mathbb{E}(N_n(k))} \to 0 \tag{6.3}$$

as $n \to \infty$. Note that the vector $(k\operatorname{avg}(\mathbf{W}_\lambda), k\operatorname{avg}(\mathbf{W}_{\lambda_0}))$ has a bivariate normal distribution with variance one and correlation st/k^2 where s is the number of common rows between λ, λ_0 and t is the number of common columns between λ, λ_0.

Define,

$$E(s,t) := \binom{n}{k}^{-2}\binom{k}{s}\binom{n-k}{k-s}\binom{k}{t}\binom{n-k}{k-t}\frac{\mathbb{P}(Z_{st} \geq k\tau \mid Z \geq k\tau)}{\bar{\Phi}(k\tau)}$$

for $1 \leq s,t \leq k$ where $Z_{st} = k^{-2}stZ + \sqrt{1-k^{-4}s^2t^2}Z'$ and Z, Z' are i.i.d. standard Gaussian r.v.s. Thus we have

$$I_n := \sum_{s=1}^{k}\sum_{t=1}^{k} E(s,t) \tag{6.4}$$

Clearly $E(k,k) = 1/\mathbb{E}(N_n(k)) \to 0$ as $n \to \infty$. We need to estimate $E(s,t)$ for $st \leq k(k-1)$.

Using Lemma 5.4 and Lemma 5.1 with $\theta_{st} := \sqrt{\frac{k^2-st}{k^2+st}}$, we have

$$\frac{\mathbb{P}(Z_{st} \geq k\tau \mid Z \geq k\tau)}{\bar{\Phi}(k\tau)} \leq \frac{2\bar{\Phi}(\theta_{st}k\tau)}{\bar{\Phi}(k\tau)} \leq 4\sqrt{\frac{k^2-st}{k^2+st}}\exp\left(\frac{rck^2\tau^2}{k^2+rc}\right)$$

for $st \leq k(k-1)$ and $k\tau > 1$. Now we use Lemma 5.6 with $N = \binom{n}{k}^2, x = (\tilde{k}-k)\tau a_N \approx k\tau^2$. Thus we have $I_n \to 0$ as $n \to \infty$ and we are done. ■

7. Proofs: Local optima

7.1. **Proof of Structure Theorem 3.7.** Let $\mathcal{R}_k$ and $\mathcal{C}_k$ be the events that the sub-matrix $\mathbf{C} = \mathbf{W}_{[k]\times[k]}$ is row optimal and is column optimal, respectively. Clearly $\mathbb{P}(\mathcal{R}_k) = \mathbb{P}(\mathcal{C}_k) = \binom{n}{k}^{-1}$ and we need the probability of the event $\mathcal{I}_k := \mathcal{R}_k \cap \mathcal{C}_k$ as well as the conditional distribution of $\mathbf{C}$ given $\mathcal{I}_k$.

Let $c_{i\cdot} = k^{-1}\sum_{j=1}^k w_{ij}, c_{\cdot j} = k^{-1}\sum_{i=1}^k w_{ij}, c_{\cdot\cdot} = k^{-2}\sum_{i,j=1}^k w_{ij}$ for $1 \leq i,j \leq n$. We use the ANOVA decomposition of the Gaussian matrix $\mathbf{C}$ which gives that

$$c_{ij} = \tilde{c}_{ij} + (c_{i\cdot} - c_{\cdot\cdot}) + (c_{\cdot j} - c_{\cdot\cdot}) + c_{\cdot\cdot}$$

where $\tilde{c}_{ij} = c_{ij} - c_{i\cdot} - c_{\cdot j} + c_{\cdot\cdot}$. Under the gaussian assumption the random variables $\tilde{\mathbf{C}} = ((\tilde{c}_{ij}))_{i,j=1}^k, (c_{i\cdot} - c_{\cdot\cdot})_{i=1}^k, (c_{\cdot j} - c_{\cdot\cdot})_{j=1}^k, c_{\cdot\cdot}$ are independent and obviously independent of the remaining $n-k$ row and column averages $(c_{i\cdot})_{i=k+1}^n, (c_{\cdot j})_{j=k+1}^n$.

Clearly we have

$$\mathcal{R}_k := \{\min_{1\leq j\leq k} c_{\cdot j} \geq \max_{k<j\leq n} c_{\cdot j}\} = \{c_{\cdot\cdot} - \max_{1\leq j\leq k}\{c_{\cdot\cdot} - c_{\cdot j}\} \geq \max_{k<j\leq n} c_{\cdot j}\}$$
$$\mathcal{C}_k := \{\min_{1\leq i\leq k} c_{i\cdot} \geq \max_{k<i\leq n} c_{i\cdot}\} = \{c_{\cdot\cdot} - \max_{1\leq i\leq k}\{c_{\cdot\cdot} - c_{i\cdot}\} \geq \max_{k<i\leq n} c_{i\cdot}\}$$

and hence

$$\begin{aligned}\mathcal{I}_k = \mathcal{R}_k \cap \mathcal{C}_k = \{c_{\cdot\cdot} \geq \max_{k<j\leq n}\{c_{\cdot j}, c_{j\cdot}\}\} \\ \cap \{\max_{1\leq j\leq k}\{c_{\cdot\cdot} - c_{\cdot j}\} \leq c_{\cdot\cdot} - \max_{k<j\leq n} c_{\cdot j}\} \\ \cap \{\max_{1\leq i\leq k}\{c_{\cdot\cdot} - c_{i\cdot}\} \leq c_{\cdot\cdot} - \max_{k<i\leq n} c_{i\cdot}\}.\end{aligned}$$

Define $M_{n-k} = a_n(\sqrt{k}\max_{k<j\le n} c_{\cdot j} - b_n), M'_{n-k} = a_n(\sqrt{k}\max_{k<i\le n} c_{i\cdot} - b_n)$ for the recentered and rescaled maxima of the remaining row and column averages. From Lemma 3.6, (M_{n-k}, M'_{n-k}) converges in distribution to $(-\log T, -\log T')$ where T, T' are i.i.d. Exponential rate one random variables. Now the event $\mathcal{I}_k$ can be expressed as

$$\begin{aligned}\mathcal{I}_k = \{\sqrt{k}c_{\cdot\cdot} \ge b_n + a_n^{-1}\max\{M_{n-k}, M'_{n-k}\}\} \\ \cap\{\sqrt{k}\max_{1\le j\le k}\{c_{\cdot\cdot} - c_{\cdot j}\} \le \sqrt{k}c_{\cdot\cdot} - (b_n + a_n^{-1}M_{n-k})\} \\ \cap\{\sqrt{k}\max_{1\le i\le k}\{c_{\cdot\cdot} - c_{i\cdot}\} \le \sqrt{k}c_{\cdot\cdot} - (b_n + a_n^{-1}M'_{n-k})\}.\end{aligned} \tag{7.1}$$

Note that $kc_{\cdot\cdot}, \sqrt{k}c_{i\cdot}, \sqrt{k}c_{\cdot j}$ are standard Gaussian random variables. Let $\bar{\Phi}(x) = \mathbb{P}(Z \ge x)$ be the tail probability of standard Gaussian distribution. Define $F_k(x) = \mathbb{P}(\max_{1\le i\le k}(\bar{Z} - Z_i) \le x)$ where $Z_1, Z_2, \ldots, Z_k$ are i.i.d. standard Gaussian r.v.s with $\bar{Z} = k^{-1}\sum_{i=1}^k Z_i$.

Using the independence resulting from the ANOVA Decomposition and using (7.1) gives

$$\mathbb{P}(\mathcal{I}_k) = \mathbb{E}\big(F_k(k^{-1/2}Z - b_n - a_n^{-1}M_{n-k}) \tag{7.2}$$

$$F_k(k^{-1/2}Z - b_n - a_n^{-1}M'_{n-k}) \tag{7.3}$$

$$\mathbb{1}\{k^{-1/2}Z \ge b_n + a_n^{-1}\max\{M_{n-k}, M'_{n-k}\}\}\big). \tag{7.4}$$

The idea of the rest of the proof is as follows. We will show that the event (7.4) has probability of the order $(\log n)^{(k-1)/2}n^{-k}$ and conditionally on this event $(a_n(k^{-1/2}Z - b_n), M_{n-k}, M'_{n-k})$ converges in distribution. In particular $k^{-1/2}Z - b_n - M_{n-k}/a_n = \Theta_P(a_n^{-1})$. Using Lemma 5.3(a) we get that the first two random variables namely (7.2) and (7.3) behaves like $a_n^{-(k-1)} \approx (\log n)^{-(k-1)/2}$. Combining everything we will finally get the order of the probability to be $(\log n)^{-(k-1)/2}n^{-k}$.

Using Lemma 5.2 with $\theta = k$ we have that

$$\sqrt{k}(\sqrt{2\pi}b_n)^{1-k}n^k\,\mathbb{P}(k^{-1/2}Z \ge b_n + a_n^{-1}x) \to e^{-kx}$$

as $n \to \infty$ uniformly over $|x| \ll a_n$. Using the fact that $\max\{M_{n-k}, M'_{n-k}\}$ converges in distribution to $-\log(T/2)$ where $T \sim \text{Exp}(1)$ we have

$$\begin{aligned}\sqrt{k}(\sqrt{2\pi}b_n)^{1-k}n^k\,\mathbb{P}(k^{-1/2}Z \ge b_n + a_n^{-1}\max\{M_{n-k}, M'_{n-k}\}) \\ \to 2^{-k}\,\mathbb{E}(T^k) = 2^{-k}k!.\end{aligned}$$

Now for any $x, x', y < \min\{x, x'\}$ fixed, we have

$$\begin{aligned}\sqrt{k}(\sqrt{2\pi}b_n)^{1-k}n^k\,\mathbb{P}(k^{-1/2}Z \ge b_n - a_n^{-1}\log y, M_{n-k} \le -\log x, M'_{n-k} \le -\log x') \\ \to y^k e^{-x-x'} \text{ as } n \to \infty.\end{aligned}$$

From here it easily follows that, conditional on $\{k^{-1/2}Z \ge b_n + a_n^{-1}\max\{M_{n-k}, M'_{n-k}\}\}$,

$$(a_n(k^{-1/2}Z - b_n), M_{n-k}, M'_{n-k}) \Rightarrow (-\log G, -\log(G+Y), -\log(G+Y')) \tag{7.5}$$

where Y, Y' are i.i.d. Exp(1) and $G \sim \text{Gamma}(k, 2)$ with density $\frac{2^k}{(k-1)!}x^{k-1}e^{-2x}, x > 0$.

Using Lemma 5.3(a) we finally have

$$\begin{aligned}&\lim_{n\to\infty}\binom{n}{k}{a_n}^{k-1}\,\mathbb{P}(\mathcal{I}_k)\\&=\frac{2^{-k}(2\pi)^{(k-1)/2}}{\sqrt{k}}\lim_{n\to\infty}\mathbb{E}({a_n}^{2(k-1)}F_k(a_n^{-1}\log(1+Y/G))F_k(a_n^{-1}\log(1+Y'/G)))\\&=\frac{(2\pi)^{(k-1)/2}f_k^2}{\sqrt{k}2^k}\,\mathbb{E}((\log(1+Y/G)\log(1+Y'/G))^{k-1}).\end{aligned}$$

In particular, we have

$$\mathbb{P}(\mathcal{I}_k)=\frac{\theta_k}{\binom{n}{k}(\log n)^{(k-1)/2}}(1+o(1))\text{ as }n\to\infty$$

where

$$\theta_k:=\frac{k^{2k+1/2}}{\pi^{(k-1)/2}2^{2k-1}k!^2}\,\mathbb{E}((\log(1+Y/G)\log(1+Y'/G))^{k-1}).\tag{7.6}$$

The above gives the asymptotic probability of $\mathbf{C}$ being a local optima. Let us now find the asymptotic conditional distribution of the matrix itself given $\mathcal{I}_k$. In matrix form, the ANOVA decomposition can be written as

$$\mathbf{C}=c_{\cdot\cdot}\mathbf{1}\mathbf{1}'+\tilde{\mathbf{C}}+\begin{bmatrix}c_{1\cdot}-c_{\cdot\cdot}\\c_{2\cdot}-c_{\cdot\cdot}\\\vdots\\c_{k\cdot}-c_{\cdot\cdot}\end{bmatrix}\mathbf{1}'+\mathbf{1}\begin{bmatrix}c_{\cdot1}-c_{\cdot\cdot}\\c_{\cdot2}-c_{\cdot\cdot}\\\vdots\\c_{\cdot k}-c_{\cdot\cdot}\end{bmatrix}'\tag{7.7}$$

where $\tilde{\mathbf{C}}=(\tilde{c}_{ij})$ where $\tilde{c}_{ij}=c_{ij}-c_{i\cdot}-c_{\cdot j}+c_{\cdot\cdot},1\le i,j\le k$ is independent of the event $\mathcal{I}_k$. This immediately gives the second term $\tilde{\mathbf{Z}}$ in the structure theorem. Now note that by (7.1) on $\mathcal{I}_k$ the row sums satisfy

$$\sqrt{k}\max_{1\le j\le k}\{c_{\cdot\cdot}-c_{\cdot j}\}\le\sqrt{k}c_{\cdot\cdot}-(b_n+a_n^{-1}M_{n-k})$$

Here the term on the left has distribution $\max_{1\leqslant j\leqslant k}\left\{\bar{Z}-Z_i\right\}$ where the Z_i are i.i.d. standard gaussian random variables and $\bar{Z}=\operatorname{avg}(\{Z_i\}_{1\leqslant i\leqslant k})$. Further by the ANOVA decomposition, this random variable is independent of the term on the right which by (7.5) is of order $\Theta_P(a_n^{-1})$. In fact, (7.5) implies that conditional on the event $\{k^{1/2}c_{\cdot\cdot}\ge b_n+a_n^{-1}\max\{M_{n-k},M'_{n-k}\}\}$, the random variable $a_n(\sqrt{k}c_{\cdot\cdot}-(b_n+a_n^{-1}M_{n-k}))$ converges in distribution to $\log(1+Y/G)$. Thus for the third term in the ANOVA decomposition in (7.7), on the event $\mathcal{I}_k$ intuitively one is looking at the distribution of $(Z_1-\bar{Z},Z_2-\bar{Z},\ldots,Z_k-\bar{Z})$ conditional on $\max_{1\leqslant j\leqslant k}\left\{\bar{Z}-Z_i\right\}\leqslant\Theta_P(a_n^{-1})$ which is exactly the type of event Lemma 5.3(b) is geared to tackle. An identical argument applies to the last term in (7.7).

More precisely, using Lemma 5.3(b), (7.5) and the form of $\mathcal{I}_k$ in (7.1), a simple conditioning argument shows that the conditional distribution of $a_n(c_{\cdot\cdot}-b_n/\sqrt{k},c_{1\cdot}-c_{\cdot\cdot},\ldots,c_{k\cdot}-c_{\cdot\cdot},c_{\cdot1}-c_{\cdot\cdot},c_{\cdot k}-c_{\cdot\cdot})$ converges in distribution to $k^{-1/2}(-\log G,(kU_1-1)\log(1+T/G),\ldots,(kU_k-1)\log(1+T/G),(kU'_1-1)\log(1+T'/G),(kU'_k-1)\log(1+T'/G))$ where $(U_1,U_2,\ldots,U_k)$, $(U'_1,U'_2,\ldots,U'_k)$ are i.i.d. from Dirichlet$(1,1,\ldots,1)$ distribution and (G,T,T') has joint density

$$\propto(\log(1+t/g)\log(1+t'/g))^{k-1}g^{k-1}e^{-t-t'-2g},\ g,t,t'\ge0.$$

The $(\log(1+t/g)\log(1+t'/g))^{k-1}$ term is arising from the $(k-1)$-dimensional volume of the simplexes $\{\max_{1\le j\le k}\{\bar{x}-x_j\}\le \log(1+t/g)\}$ and $\{\max_{1\le j\le k}\{\bar{x}-x_j\}\le \log(1+t'/g)\}$. ■

7.2. **Variance asymptotics.** The previous section gave amongst other things, asymptotics for the expected number of local optima $\mathbb{E}(L_n(k))$. The aim of this Section is to prove Theorem 3.9, giving asymptotics for variance of $L_n(k)$. The first step is to understand the joint distribution of two matrices being locally optimal, which turns out to have a highly non-trivial structure. We start with a proof of Lemma 3.11 which bounds the probability for two *overlapping* matrices being locally optimal.

7.2.1. *Proof of Lemma* 3.11*.* Recall that this Lemma gives asymptotic bounds for the probability of the event $\mathcal{B}_{s,t,k}$ (see Figure 3.1) for two matrices having $k-s$ rows and $k-t$ columns in common to be both locally optimal. We define the following matrices

$$\begin{aligned} &\mathbf{X}_1=\mathbf{W}_{[s]\times[t]}, && \mathbf{X}_2=\mathbf{W}_{[s]\times[t+1,k]}, && \\ &\mathbf{X}_3=\mathbf{W}_{[s+1,k]\times[t]}, && \mathbf{X}_4=\mathbf{W}_{[s+1,k]\times[t+1,k]}, && \mathbf{X}_5=\mathbf{W}_{[s+1,k]\times[k+1,k+t]} \\ & && \mathbf{X}_6=\mathbf{W}_{[k+1,k+s]\times[t+1,k]}, && \mathbf{X}_7=\mathbf{W}_{[k+1,k+s]\times[k+1,k+t]}. \end{aligned}$$

Let $S_i=\operatorname{avg}(\mathbf{X}_i)$ and $\theta_i=$ number of entries in $\mathbf{X}_i$ for $i=1,2,\ldots,7$. In particular we have

$$\begin{aligned} &\theta_1=\theta_7=st, && \theta_2=\theta_6=s(k-t),\\ &\theta_3=\theta_5=(k-s)t \quad \text{and} && \theta_4=(k-s)(k-t). \end{aligned}$$

Clearly the joint density of $(S_i, 1\le i\le 7)$ is given by

$$\prod_{i=1}^{7}\sqrt{\theta_i/2\pi}\exp(-\theta_i s_i^2/2) \text{ for } (s_1,s_2,\ldots,s_7)\in\mathbb{R}^7.$$

Define

$$M_c=\max_{k+t+1\le j\le n}\sum_{i=1}^{k} w_{i,j} \text{ and } M_c'=\max_{k+t+1\le j\le n}\sum_{i=1}^{k} w_{s+i,j},$$

i.e., M_c is the maximum column sum of the sub-matrix $\mathbf{W}_{[k]\times[k+t+1,n]}$ and M_c' is the maximum column sum of the sub-matrix $\mathbf{W}_{[s+1,s+k]\times[k+t+1,n]}$. Similarly define

$$M_r=\max_{k+s+1\le i\le n}\sum_{j=1}^{k} w_{i,j} \text{ and } M_r'=\max_{k+s+1\le i\le n}\sum_{j=1}^{k} w_{t+i,j}$$

as the maximum row sum of the sub-matrix $\mathbf{W}_{[k+s+1,n]\times[k]}$ and the maximum row sum of the sub-matrix $\mathbf{W}_{[k+s+1,n]\times[t+1,t+k]}$, respectively. For a real number $x\in\mathbb{R}$, let $\mathcal{D}(x)$ be the set

$$\begin{aligned} \mathcal{D}(x)=\{(s_1,s_2,\ldots,s_7)\in\mathbb{R}^7 \mid\ & ts_1+(k-t)s_2\ge x, && ss_1+(k-s)s_3\ge x,\\ & ss_2+(k-s)s_4\ge x, && ts_3+(k-t)s_4\ge x,\\ & (k-t)s_4+ts_5\ge x, && (k-s)s_4+ss_6\ge x,\\ & (k-s)s_5+ss_7\ge x, && (k-t)s_6+ts_7\ge x\}. \end{aligned}$$

Note that $\mathcal{D}(x)$ is decreasing in x. It is easy to see that

$$\mathcal{B}_{s,t,k}\subseteq\{(S_1,S_2,\ldots,S_7)\in\mathcal{D}(\min\{M_r,M_r',M_c,M_c'\})\}.$$

Now (M_r, M_r', M_c, M_c') is independent of $(S_i, 1 \le i \le 7)$ and $M_r \overset{\mathrm{d}}{=} M_r', M_c \overset{\mathrm{d}}{=} M_c'$. Thus we have

$$\mathbb{P}(\mathcal{B}_{s,t,k}) \le 2\,\mathbb{E}(f(M_r) + f(M_c))$$

where $f(x) := \mathbb{P}((S_1, S_2, \dots, S_7) \in \mathcal{D}(x))$. We claim that

$$f(x) \le \exp\left(-\left(1 - \frac{(k-s)(k-t)}{2k^2 - st}\right) x^2\right) \tag{7.8}$$

for $x > 0$. Using standard calculus one can easily check that under $\mathcal{D}(x)$, $\sum_{i=1}^7 \theta_i s_i^2$ is minimized at $s_i = a_i, 1 \le i \le 7$ where

$$a_1 = a_7 = \frac{(3k-s-t)x}{2k^2 - st}, \qquad a_2 = a_6 = \frac{(2k-t)x}{2k^2 - st},$$
$$a_3 = a_5 = \frac{(2k-s)x}{2k^2 - st} \text{ and } a_4 = \frac{2kx}{2k^2 - st}.$$

We will not use this fact in the subsequent calculations, however it will help us to estimate $\mathbb{P}(\mathcal{D}(x))$. Note that the above vector $(a_1, a_2, \dots, a_7)$ makes all the inequalities in $\mathcal{D}(x)$ equalities. We have

$$\begin{aligned} f(x) &:= \mathbb{P}((S_1, S_2, \dots, S_7) \in \mathcal{D}(x)) \\ &= \int_{\mathcal{D}(x)} \prod_{i=1}^7 \sqrt{\theta_i/2\pi}\exp(-\theta_i s_i^2/2) ds_i \\ &= \int_{\mathcal{D}(0)} \prod_{i=1}^7 \sqrt{\theta_i/2\pi}\exp(-\theta_i (s_i + a_i)^2/2) ds_i \\ &= \prod_{i=1}^7 \sqrt{\theta_i/2\pi}\exp(-\theta_i a_i^2/2) \int_{\mathcal{D}(0)} \exp(-\sum_{i=1}^7 \theta_i(s_i^2/2 + a_i s_i)) ds_i. \end{aligned}$$

Further note that

$$\begin{aligned} \frac{2k^2 - st}{x}\sum_{i=1}^7 \theta_i a_i s_i =& s(2k-s-t)(ts_1 + (k-t)s_2 + (k-t)s_6 + ts_7) \\ &+ kt(ss_1 + (k-s)s_3 + (k-s)s_5 + ss_7) \\ &+ s(k-t)(ss_2 + (k-s)s_4 + (k-s)s_4 + ss_6) \\ &+ (k-s)(k-t)(ts_3 + (k-t)s_4 + (k-t)s_4 + ts_5) \end{aligned}$$

which is non-negative under $\mathcal{D}(0)$. Thus we have

$$\begin{aligned} f(x) &\le \prod_{i=1}^7 \sqrt{\theta_i/2\pi}\exp(-\theta_i a_i^2/2) \int_{\mathcal{D}(0)} \exp(-\sum_{i=1}^7 \theta_i s_i^2/2) ds_i \\ &\le \exp(-\sum_{i=1}^7 \theta_i a_i^2/2). \end{aligned}$$

Simplifying we have

$$\frac{1}{2}\sum_{i=1}^7 \theta_i a_i^2 = \left(1 - \frac{(k-s)(k-t)}{2k^2 - st}\right) x^2.$$

This proves the claim (7.8). Note that $M_r \overset{\mathrm{d}}{=} \sqrt{k} V_{n-k-s}$ and $M_c \overset{\mathrm{d}}{=} \sqrt{k} V_{n-k-t}$ where $V_n = \max\{Z_1, Z_2, \ldots, Z_n\}$ and Z_i's are i.i.d. $\mathrm{N}(0,1)$. Thus to complete the proof it is enough to show that for any constant $\theta > 0$ we have

$$\mathbb{E}(\exp(-\theta \max\{V_n, 0\}^2)) \le \eta(\theta) \exp(-\theta b_n^2)$$

for some constant $\eta(\theta) > 0$ where b_n satisfies $e^{-b_n^2/2} = \sqrt{2\pi} b_n / n$. Using $\theta = (1 - (k-s)(k-t)/(2k^2 - st))$ and the fact that $b_n = \sqrt{2 \log n} - \log(4\pi \log n)/\sqrt{8 \log n}$ gives the bound asserted for $\mathbb{P}(\mathcal{B}_{s,t,k})$. The following Lemma 7.1 completes the proof. ■

Lemma 7.1. *Let $V_n := \max\{Z_1, Z_2, \ldots, Z_n\}$ where Z_i's are i.i.d. $N(0,1)$. For any constant $\theta > 0$ we have*

$$\mathbb{E}(\exp(-\theta \max\{V_n, 0\}^2)) \le \eta(\theta) \exp(-\theta b_n^2)$$

for some constant $\eta(\theta) > 0$ for all $n \ge 1$ where $b_n = \sqrt{2 \log n} - \log(4\pi \log n)/\sqrt{8 \log n}$.

Proof of Lemma 7.1. Define $X_n = b_n(b_n - V_n)$. We have

$$\mathbb{E}(\exp(-\theta \max\{V_n, 0\}^2)) \le 2^{-n} + \exp(-\theta b_n^2)\, \mathbb{E}(\exp(2\theta X_n) \mathbb{1}\{V_n \ge 0\}).$$

It is easy to see that $2^{-n} \exp(\theta b_n^2)$ is bounded for any n. Thus we have to show that $\mathbb{E}(\exp(2\theta X_n) \mathbb{1}\{V_n \ge 0\})$ is uniformly bounded in n. By Lemma 3.6, we have $X_n \overset{\mathrm{d}}{\Longrightarrow} \log T$ where $T \sim \exp(1)$ and $\mathbb{E}(T^{2\theta}) < \infty$. Thus $\mathbb{E}(\exp(\theta X_n) \mathbb{1}\{X_n \le c\})$ is uniformly bounded over n for any fixed $c > 0$. Moreover $V_n \ge 0$ implies $X_n \le b_n^2$. Thus we need to bound

$$\mathbb{E}(\exp(\theta X_n) \mathbb{1}\{c \le X_n \le b_n^2\})$$

for an appropriate choice of c. We will break the interval $[c, b_n^2]$ into several subintervals and estimate the contribution from each subintervals. Note that

$$\mathbb{P}(X_n \ge x) = (1 - \bar{\Phi}(b_n - x/b_n))^n \le \exp(-n\bar{\Phi}(b_n - x/b_n)).$$

Moreover, we have for all $x \in [0, b_n]$

$$\frac{b_n(b_n - x/b_n)}{1 + (b_n - x/b_n)^2} e^{-x^2/2b_n^2} \ge C$$

for some constant $C \in (0, 2\theta/e]$. Thus using the bound $\bar{\Phi}(u) \ge u^2/(\sqrt{2\pi}(1+u^2)) e^{-u^2/2}$ from Lemma 5.1 and the fact that $n e^{-b_n^2/2}/(\sqrt{2\pi} b_n) = 1 + o(1)$ we have

$$\mathbb{P}(X_n \ge x) \le \exp(-Ce^x)$$

for all $x \in [0, b_n]$.

We take $x_0 = b_n^2$ and $x_{i+1} = \log(2\theta x_i / C)$ for $i \ge 0$. Note that $2\theta/C \ge e$. Let $c_* \ge 1$ be the largest solution to the equation $x = \log(2\theta x/C)$. It is easy to see that $x_i \to c_*$ as $i \to \infty$. Thus there exists k such that $c^* < x_{k+1} < 2c_* \le x_k$. Note that $x_i \ge C x_{i+1}^2 / 4\theta$

implies that $x_i \geq 4\theta(Cx_{k+1}/4\theta)^{2^{k+1-i}}/C$. Thus we have

$$\begin{aligned}
\mathbb{E}(\exp(\theta X_n)\mathbb{1}\{2c^* \leq X_n \leq b_n^2\}) &\leq \sum_{i=0}^{k} \mathbb{E}(\exp(\theta X_n)\mathbb{1}\{x_{i+1} \leq X_n \leq x_i\}) \\
&\leq \sum_{i=0}^{k} \exp(\theta x_i)\,\mathbb{P}(X_n \geq x_{i+1}) \\
&\leq \sum_{i=0}^{k} \exp(\theta x_i - Ce^{x_{i+1}}) \leq \sum_{i=0}^{k} \exp(-\theta x_i) = O(1).
\end{aligned}$$

This completes the proof. ■

7.2.2. *Proof of Variance asymptotics: Theorem* 3.9. Let us now complete the proof of the main result.

Proof. Let $p_n = \mathbb{P}(\mathbf{W}_{[k]\times[k]}$ is locally optimal as a sub matrix of $\mathbf{W}_{[n]\times[n]})$. By Theorem 3.7 we have $p_n = (1+o(1))\theta_k/(\binom{n}{k}(\log n)^{(k-1)/2})$. By symmetry we have

$$\begin{aligned}
&\mathrm{Var}(L_n(k)) \\
&= \sum_{\lambda,\gamma\in\mathscr{S}_n(k)} \mathrm{Cov}(\mathbb{1}\{\mathbf{W}_\lambda \text{ is locally optimal}\}, \mathbb{1}\{\mathbf{W}_\gamma \text{ is locally optimal}\}) \\
&= \binom{n}{k}^2 \sum_{s=0}^{k}\sum_{t=0}^{k} \binom{k}{s}\binom{k}{t}\binom{n-k}{s}\binom{n-k}{t} \mathrm{Cov}(\mathbb{1}\{\mathbf{W}_{[k]\times[k]} \text{ is locally optimal}\}, \\
&\qquad\qquad \mathbb{1}\{\mathbf{W}_{[s+1,s+k]\times[t+1,t+k]} \text{ is locally optimal}\}).
\end{aligned}$$

Define

$$\begin{aligned}
v_n(s,t) := \binom{n}{k}^2 \binom{k}{s}\binom{k}{t}\binom{n-k}{s}\binom{n-k}{t} &\mathrm{Cov}(\mathbb{1}\{\mathbf{W}_{[k]\times[k]} \text{ is locally optimal}\}, \\
&\mathbb{1}\{\mathbf{W}_{[s+1,s+k]\times[t+1,t+k]} \text{ is locally optimal}\}) \qquad (7.9)
\end{aligned}$$

for $0 \leq s,t \leq k$. We will show the dominant contribution comes from the $s = t = k$ case, i.e. when the matrices are completely disjoint, sharing no rows and columns and in this case $v_n(k,k) \approx n^{2k^2/(k+1)} = n^{2k-2+2/(k+1)}$ with logarithmic corrections. We consider several different cases depending on the values of s,t.

Case 1. $s = t = 0$: The matrices are the same. Clearly here $0 < v_n(0,0) = \binom{n}{k}^2 p_n(1-p_n) = O((\log n)^{-(k-1)/2})$.

Case 2. $s = 0, t > 0$ or $s > 0, t = 0$: Here the matrices have identical row or column sets but do not overlap. In this case obviously both matrices cannot be simultaneously locally optimal. Thus the covariance term is $-p_n^2$ and the contribution is $|v_n(s,t)| = O(n^{2k+s+t}p_n^2) = O(n^k(\log n)^{1-k})$.

Case 3. $0 < s,t < k$: Using Lemma 3.11 we have

$$\begin{aligned}0 &\le \operatorname{Cov}(\mathbb{1}\{\mathbf{W}_{[k]\times[k]} \text{ is locally optimal}\}, \mathbb{1}\{\mathbf{W}_{[s+1,s+k]\times[t+1,t+k]} \text{ is locally optimal}\})\\ &\le \eta(s,t,k)(\log n/n^2)^{k-k(k-s)(k-t)/(2k^2-st)}.\end{aligned}$$

Thus we have

$$0 \le v_n(s,t) = O(n^{s+t+2k(k-s)(k-t)/(2k^2-st)}(\log n)^{k-k(k-s)(k-t)/(2k^2-st)})$$

Note that

$$\frac{2k(k-s)(k-t)}{2k^2-st} = \frac{2k}{\frac{k(3k-s-t)}{(k-s)(k-t)}-1} \le \frac{2k}{\frac{k(3k-s-t)}{(k-(s+t)/2)^2}-1}.$$

Thus defining $\theta := (s+t)/2k$ we have

$$s+t+\frac{2k(k-s)(k-t)}{2k^2-st} \le 2k\left(\theta+\frac{(1-\theta)^2}{2-\theta^2}\right) = 2k\left(\frac{3-\theta^3}{2-\theta^2}-1\right).$$

Now $\frac{d}{d\theta}\frac{3-\theta^3}{2-\theta^2} = \frac{6(1-\theta)+\theta^3}{(2-\theta^2)^2} > 0$ implies that $\frac{3-\theta^3}{2-\theta^2}$ is a strictly increasing function of θ. Also note that $\theta \in [1/k, 1-1/k]$. Thus we have $0 < v_n(s,t) \le v_n(k-1,k-1)$ for all $s,t \in \{1,2,\ldots,k-1\}$. Now for $s=t=k-1$ we have

$$\begin{aligned}s+t+\frac{2k(k-s)(k-t)}{2k^2-st} &= 2k-2+\frac{2k}{k^2+2k-1}\\ &= \frac{2k^2}{k+1}-\frac{2(k-1)}{(k+1)(k^2+2k-1)}.\end{aligned}$$

Thus for $0 < s,t < k$ we have

$$0 \le v_n(s,t) = O(n^{\frac{2k^2}{k+1}-\frac{2(k-1)}{(k+1)(k^2+2k-1)}}(\log n)^{k-\frac{k}{k^2+2k-1}}).$$

Case 4. $s=t=k$: Here we will show that

$$v_n(k,k) = (\nu_k+o(1))n^{2k^2/(k+1)}(\log n)^{-k^2/(k+1)}$$

for some constant $\nu_k > 0$. Let $\mathcal{I}_{k,n-k}$ be the event that $\mathbf{W}_{[k]\times[k]}$ is locally optimal as a submatrix of $\mathbf{W}_{[k]\cup[2k+1,n]\times[k]\cup[2k+1,n]}$ and let $\mathcal{I}'_{k,n-k}$ be the event that $\mathbf{W}_{[k+1,2k]\times[k+1,2k]}$ is locally optimal as a submatrix of $\mathbf{W}_{[k+1,n]\times[k+1,n]}$. Clearly these two events are independent and $\mathbb{P}(\mathcal{I}_{k,n-k}) = \mathbb{P}(\mathcal{I}'_{k,n-k}) = p_{n-k}$. Let $\mathbf{W}^* = (w^*_{ij})_{k\times k}, \mathbf{W}^{**} = (w^{**}_{ij})_{k\times k}$ denote the matrices $\mathbf{W}_{[k]\times[k]}, \mathbf{W}_{[k+1,2k]\times[k+1,2k]}$ respectively conditional on $\mathcal{I}_{k,n-k}\cap\mathcal{I}'_{k,n-k}$. From Theorem 3.7 we know refined asymptotics for this distribution. Write $\mathbf{C} := \mathbf{W}_{[k]\times[k+1,2k]}, \mathbf{C}' := \mathbf{W}_{[k+1,2k]\times[k]}$ for the remaining two submatrices that determine the local optimality of $\mathbf{W}_{[k]\times[k]}, \mathbf{W}_{[k+1,2k]\times[k+1,2k]}$. See figure 7.2.2 for a pictorial description.

By conditioning first on $\mathcal{I}_{n-k}\cap\mathcal{I}'_{n-k}$ we have

$$\begin{aligned}&\mathbb{P}(\mathbf{W}_{[k]\times[k]} \ \&\ \mathbf{W}_{[k+1,2k]\times[k+1,2k]} \text{ are locally optimal})\\ &= p^2_{n-k}\,\mathbb{E}\Big(\mathbb{1}\{\min w^*_{i\cdot} \ge \max c'_{i\cdot}, \min w^*_{\cdot j} \ge \max c_{\cdot j}\}\mathbb{1}\{\min w^{**}_{i\cdot} \ge \max c_{i\cdot}, \min w^{**}_{\cdot j} \ge \max c'_{\cdot j}\}\Big)\end{aligned}$$

Now by the ANOVA decomposition, note that $\max_i c_{i\cdot} - c_{\cdot\cdot}, \max_j c_{\cdot j} - c_{\cdot\cdot}, \max_i c'_{i\cdot} - c'_{\cdot\cdot}, \max_j c'_{\cdot j} - c'_{\cdot\cdot}, c_{\cdot\cdot}, c'_{\cdot\cdot}$ are independent. Let $d_{\cdot\cdot}, d'_{\cdot\cdot}$ be i.i.d. copies of $c_{\cdot\cdot}, c'_{\cdot\cdot}$. Then

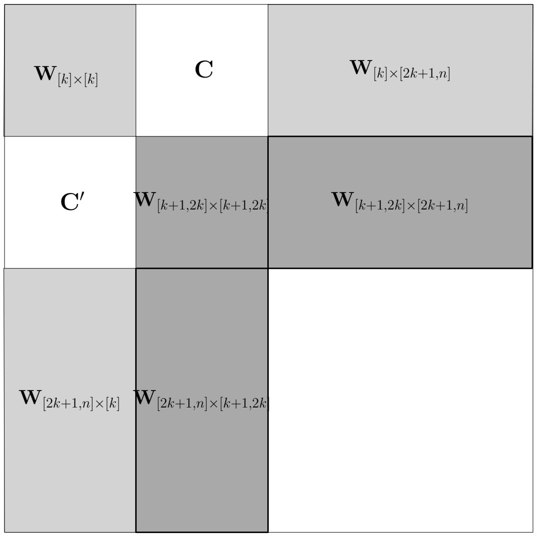
$\mathbf{W}_{[k]\times[k]}$
$\mathbf{C}$
$\mathbf{W}_{[k]\times[2k+1,n]}$
$\mathbf{C}'$
$\mathbf{W}_{[k+1,2k]\times[k+1,2k}$
$\mathbf{W}_{[k+1,2k]\times[2k+1,n]}$
$\mathbf{W}_{[2k+1,n]\times[k]}$
$\mathbf{W}_{[2k+1,n]\times[k+1,2k}$

FIGURE 7.1. The event $\mathcal{I}_{k,n-k}$ corresponds to the matrix $\mathbf{W}_{[k]\times[k]}$ being optimal in the light gray region and the event $\mathcal{I}'_{k,n-k}$ corresponds to the matrix $\mathbf{W}_{[k+1,2k]\times[k+1,2k]}$ being optimal in the dark gray region.

$(\max c'_{i\cdot}, \max c_{\cdot j})$ is independent of $(d_{\cdot\cdot} + \max c_{i\cdot} - c_{\cdot\cdot}, \max c'_{\cdot j} - c'_{\cdot\cdot} + d'_{\cdot\cdot}) \overset{d}{=} (\max c_{i\cdot}, \max c'_{\cdot j})$. This implies

$$\begin{aligned} p_n^2 = p_{n-k}^2\, \mathbb{E}\big(&\mathbb{1}\{\, \min w^*_{i\cdot} \geq \max c'_{i\cdot}, \min w^*_{\cdot j} \geq \max c_{\cdot j}\} \\ &\mathbb{1}\{\min w^{**}_{i\cdot} \geq \max c_{i\cdot} - c_{\cdot\cdot} + d_{\cdot\cdot}, \min w^{**}_{\cdot j} \geq \max c'_{\cdot j} - c'_{\cdot\cdot} + d'_{\cdot\cdot}\}\big). \end{aligned}$$

Putting these equations together results in

$$\begin{aligned} &\mathrm{Cov}(\mathbb{1}\{\mathbf{W}_{[k]\times[k]} \text{ is locally optimal}\}, \mathbb{1}\{\mathbf{W}_{[k+1,2k]\times[k+1,2k]} \text{ is locally optimal}\}) \\ &= p_{n-k}^2\, \mathbb{E}\Bigg(\mathbb{1}\{\min w^*_{i\cdot} \geq \max c'_{i\cdot}, \min w^*_{\cdot j} \geq \max c_{\cdot j}\} \\ &\qquad\qquad \cdot \Bigg(\mathbb{1}\{\min w^{**}_{i\cdot} \geq \max c_{i\cdot}, \min w^{**}_{\cdot j} \geq \max c'_{\cdot j}\} \\ &\qquad - \mathbb{1}\{\min w^{**}_{i\cdot} \geq \max c_{i\cdot} - c_{\cdot\cdot} + d_{\cdot\cdot}, \min w^{**}_{\cdot j} \geq \max c'_{\cdot j} - c'_{\cdot\cdot} + d'_{\cdot\cdot}\}\Bigg)\Bigg). \end{aligned}$$

Define the random variables

$$\begin{aligned} E &:= \min w^*_{i\cdot} - \max(c'_{i\cdot} - c'_{\cdot\cdot}), \qquad & F &:= \min w^*_{\cdot j} - \max(c_{\cdot j} - c_{\cdot\cdot}), \\ G &:= \min w^{**}_{i\cdot} - \max(c_{i\cdot} - c_{\cdot\cdot}), \qquad & H &:= \min w^{**}_{\cdot j} - \max(c'_{\cdot j} - c'_{\cdot\cdot}). \end{aligned}$$

Note that the random variables E, F, G, H are independent of $c_{..}, c'_{..}, d_{..}, d'_{..}$. Using the fact that by construction $c_{..} \overset{d}{=} d_{..}$ and $c'_{..} \overset{d}{=} d'_{..}$, we have

$$\begin{aligned}
&p_{n-k}^{-2} \operatorname{Cov}(\mathbb{1}\{\mathbf{W}_{[k]\times[k]} \text{ is locally optimal}\}, \mathbb{1}\{\mathbf{W}_{[k+1,2k]\times[k+1,2k]} \text{ is locally optimal}\}) \\
&\quad = \mathbb{E}(\mathbb{P}(c_{..} \le \min\{F, G\})\, \mathbb{P}(c'_{..} \le \min\{E, H\}) \\
&\qquad\quad - \mathbb{P}(c_{..} \le F)\, \mathbb{P}(c_{..} \le G)\, \mathbb{P}(c'_{..} \le E)\, \mathbb{P}(c'_{..} \le H) \mid E, F, G, H) \\
&\quad = \mathbb{E}(\mathbb{P}(c_{..} \le \min\{F, G\})\, \mathbb{P}(c'_{..} \le \min\{E, H\}) \\
&\qquad\quad (1 - \mathbb{P}(c_{..} \le \max\{F, G\})\, \mathbb{P}(c'_{..} \le \max\{E, H\})) \mid E, F, G, H) \\
&\quad = \mathbb{E}\Bigg(\mathbb{P}(c_{..} \le \min\{F, G\})\, \mathbb{P}(c'_{..} \le \min\{E, H\}) \\
&\qquad\quad \Bigg(\mathbb{P}(c_{..} \ge \max\{F, G\}) + \mathbb{P}(c_{..} \le \max\{F, G\})\, \mathbb{P}(c'_{..} \ge \max\{E, H\})\Bigg) \mid E, F, G, H\Bigg).
\end{aligned}$$

Note that from the Structure Theorem 3.7 we have that $a_n(w^{**}_{i\cdot} - b_n/\sqrt{k}), a_n(w^{*}_{\cdot j} - b_n/\sqrt{k})$ are tight. Thus we have

$$\begin{aligned}
&p_{n-k}^{-2} \operatorname{Cov}(\mathbb{1}\{\mathbf{W}_{[k]\times[k]} \text{ is locally optimal}\}, \mathbb{1}\{\mathbf{W}_{[k+1,2k]\times[k+1,2k]} \text{ is locally optimal}\}) \\
&\quad = (2 + o(1))\, \mathbb{P}(c_{..} \ge \max\{F, G\}) = (2 + o(1))\, \mathbb{P}(\max c_{\cdot j} \ge \min w^{*}_{\cdot j}, \max c_{i\cdot} \ge \min w^{**}_{i\cdot}).
\end{aligned}$$

Using Lemma 3.10 with $s = t = k, \theta = 1/\sqrt{k}$ we have

$$\begin{aligned}
v_n(k,k) &= \binom{n}{k}^2 \binom{n-k}{k}^2 p_{n-k}^2 (2 + o(1))\, \mathbb{P}(\max c_{\cdot j} \ge \min w^{*}_{\cdot j}, \max c_{i\cdot} \ge \min w^{**}_{i\cdot}) \\
&= \Theta(n^{2k - 2k/(k+1)} (\log n)^{k/(k+1) - 1 - (k-1)}) = \Theta((n/\sqrt{\log n})^{2k^2/(k+1)}).
\end{aligned}$$

Case 5. $s < k$ and $t = k$: In this case note that

$$\begin{aligned}
&\mathbb{P}(\mathbf{W}_{[k]\times[k]} \text{ is locally optimal}, \mathbf{W}_{[s+1,s+k]\times[k+1,2k]} \text{ is locally optimal}) \\
&\le \mathbb{P}(\mathbf{W}_{[k]\times[k]} \text{ is row optimal}, \mathbf{W}_{[s+1,s+k]\times[k+1,2k]} \text{ is row optimal}) = \binom{n}{k}^{-2}.
\end{aligned}$$

Thus we have $|v_n(s,k)| = O(n^{2k+s+k-2k}) = O(n^{2k-2})$ for $s \le k-2$. We need to consider the case $t = k, s = k-1$ separately as $2k - 1 > 2k^2/(k+1)$. However, using a similar analysis done in case 4 and the fact that $\mathbb{P}(\max c_{i\cdot} \ge \max w^{**}_{i\cdot}) = O(\sqrt{\log n}/n)$ where $\mathbf{C} = \mathbf{W}_{[k-1]\times[k+1,2k]}$ we have

$$|v_n(k-1,k)| = O(n^{2k+2k-1-2k-1}\sqrt{\log n}) = O(n^{2k-2}\sqrt{\log n}).$$

Note that in the case when $s = k-1, t = k$, the number of sub matrix pairs and covariance term balance each other in a subtle way.

Case 6. $s = k$ and $t < k$: Similar to Case 5.

Combining everything we finally have

$$\operatorname{Var}(L_n(k)) = (\nu_k + o(1))(n/\sqrt{\log n})^{2k^2/(k+1)}$$

for some constant $\nu_k > 0$ where the $o(1)$ term decays like

$$(\log n/n^2)^{\frac{k-1}{(k+1)(k^2+2k-1)}} (\log n)^{2k-1}. \qquad \blacksquare$$

8. Proof of the Central limit theorem

The last section analyzed first and second order properties of the number of local optima $L_n(k)$. The aim of this section is to prove the Central Limit Theorem 3.12 for $L_n(k)$, for fixed $k \geqslant 2$. For submatrix $\lambda = I \times J \in \mathscr{S}_n(k)$ define

$$\mathcal{I}_\lambda := \mathbb{1}\{\mathbf{W}_\lambda \text{ is locally optimal for } \mathbf{W}_{[n]\times[n]}\}.$$

Write $L := L_n(k) = \sum_{\lambda \in \mathscr{S}_n(k)} \mathcal{I}_\lambda$ for the total number of locally optimal sub matrices of size $k \times k$. To emphasize the dependence on the underlying matrix $\mathbf{W} := \mathbf{W}_{[n]\times[n]}$, when necessary we will write $\mathcal{I}_\lambda(\mathbf{W}), L(\mathbf{W})$ instead of $\mathcal{I}_\lambda, L$ respectively.

Let

$$p_n = \mathbb{E}(\mathcal{I}_\lambda),\ \mu = \mathbb{E}(L) = \binom{n}{k}^2 p_n \text{ and } \sigma^2 = \operatorname{Var}(L).$$

From Theorem 3.8 and Theorem 3.9 we have

$$\mu = \frac{\theta_k n^k}{k!(\log n)^{(k-1)/2}}(1+o(1)) \text{ and } \sigma^2 = \frac{\nu_k n^{2k^2/(k+1)}}{(\log n)^{k^2/(k+1)}}(1+o(1))$$

for some constant $\theta_k, \nu_k > 0$. Thus

$$\frac{\sigma}{\mu} = (1+o(1))\frac{\alpha_k}{n^{k/(k+1)}(\log n)^{1/(2k+2)}} = o(1). \tag{8.1}$$

where $\alpha_k = k!\nu_k/\theta_k > 0$. Let $\mathbf{W}' = ((w'_{ij}))$ be an i.i.d. copy of the underlying matrix $\mathbf{W}$. For any fixed submatrix $\lambda = I \times J \in \mathscr{S}_n(k)$, define

$$w^\lambda_{ab} = \begin{cases} w'_{ab} & \text{if either } a \in I \text{ or } b \in J \\ w_{ab} & \text{if } a \notin I \text{ and } b \notin J, \end{cases}$$

$\mathbf{W}^\lambda = ((w^\lambda_{ij}))$ and $L^\lambda := L(\mathbf{W}^\lambda)$. Thus we replace **all** n entries for the row set and column set of λ by independent and identical entries w^λ_{ab}. If λ is chosen uniformly at random from $\mathscr{S}_n(k)$, it is easy to see that $\mathbf{W}^\lambda$ and $\mathbf{W}$ form an exchangeable pair. However we will not use the exchangeable pair approach for Stein's method as the conditional error $\mathbb{E}(L^\lambda - L \mid \mathbf{W})$ is not linear with L. Recall from the discussion on Stein's method in Section 4.4, in order to prove that $\hat{L} = (L-\mu)/\sigma$, one needs to bound $|\,\mathbb{E}(g'(\hat{L}) - \hat{L}g(\hat{L}))|$ for g in the class of functions $\mathcal{D}'$ in (4.1). We will use a direct argument to bound this quantity.

First note that $\mathcal{I}_\lambda(\mathbf{W})$ is independent of L^λ. Thus for any twice differentiable function f, we have

$$\begin{aligned}
\mathbb{E}((L-\mu)f(L)) &= \sum_\lambda \mathbb{E}(\mathcal{I}_\lambda f(L) - p_n f(L)) \\
&= \sum_\lambda \mathbb{E}(\mathcal{I}_\lambda(f(L) - f(L^\lambda))) \\
&= \sum_\lambda \mathbb{E}\big(\mathcal{I}_\lambda((L-L^\lambda)f'(L) - \frac{1}{2}(L-L^\lambda)^2 f''(L^\lambda_*))\big)
\end{aligned}$$

where L_*^λ is a random variable. In particular with $\hat{L} = (L-\mu)/\sigma$ and $f(x) = g((x-\mu)/\sigma)$ we have

$$|\,\mathbb{E}(\hat{L}g(\hat{L}) - g'(\hat{L}))| \leq \frac{||g'||_\infty}{\sigma^2}\,\mathbb{E}\Big|\sum_\lambda \mathcal{I}_\lambda\,\mathbb{E}\big(L - L^\lambda \mid \mathbf{W}\big) - \sigma^2\Big| + \frac{||g'||_\infty}{2\sigma^3}\,\mathbb{E}\sum_\lambda \mathcal{I}_\lambda (L-L^\lambda)^2.$$

Note that by symmetry

$$\mathbb{E}\sum_\lambda \mathcal{I}_\lambda(L-L^\lambda)^2 = \mu\,\mathbb{E}((L-L^{\lambda_0})^2 \mid \mathcal{I}_{\lambda_0}) \tag{8.2}$$

where $\lambda_0 = [k]\times[k]$ and for simplicity we write $E(\cdot \mid \mathcal{I}_{\lambda_0}) := E(\cdot \mid \mathcal{I}_{\lambda_0} = 1)$. Thus using Lemma 4.1 we have

$$d_{\mathcal{W}}(\hat{L}, \mathrm{N}(0,1)) \leq \frac{1}{\sigma^2}\,\mathbb{E}\Big|\sum_\lambda \mathcal{I}_\lambda\,\mathbb{E}\big(L-L^\lambda \mid \mathbf{W}\big) - \sigma^2\Big| + \frac{\mu}{\sigma^3}\,\mathbb{E}((L-L^{\lambda_0})^2 \mid \mathcal{I}_{\lambda_0}). \tag{8.3}$$

Recall that, for $\lambda, \gamma \in \mathscr{S}_n(k)$, $|\lambda\cap\gamma| = (s,t)$ implies that λ and γ share s many rows and t many columns. For fixed $\lambda \in \mathscr{S}_n(k)$, define

$$\mathscr{S}_\lambda(s,t) := \{\gamma \in \mathscr{S}_n(k) \mid |\lambda\cap\gamma| = (k-s, k-t)\}, \qquad 0 \leq s,t \leq k.$$

Thus $\mathscr{S}_\lambda(s,t)$ consists of the set of submatrices which s rows and t columns **different** from λ. Write

$$S_\lambda(s,t) := \sum_{\gamma\in\mathscr{S}_\lambda(s,t)} (\mathcal{I}_\gamma - \mathcal{I}_\gamma(\mathbf{W}^\lambda))$$

so that we have $L - L^\lambda = \sum_{0\leq s,t\leq k} S_\lambda(s,t)$. Clearly

$$|\mathscr{S}_\lambda(s,t)| = \binom{k}{s}\binom{k}{t}\binom{n-k}{s}\binom{n-k}{t} = O(n^{s+t}).$$

Let

$$u_n(s,t) := \mathbb{E}(S_\lambda(s,t) \mid \mathcal{I}_\lambda)$$

By symmetry, this term is the same for all λ. Recall from (7.9) that the variance of $L_n(k)$ could be expressed as $\sigma^2 = \sum_{s,t} v_n(s,t)$ where

$$v_n(s,t) := \binom{n}{k}^2\binom{k}{s}\binom{k}{t}\binom{n-k}{s}\binom{n-k}{t}\,\mathrm{Cov}(\mathbb{1}\{\mathbf{W}_{[k]\times[k]} \text{ is locally optimal}\},$$
$$\mathbb{1}\{\mathbf{W}_{[s+1,s+k]\times[t+1,t+k]} \text{ is locally optimal}\})$$

A simple conditioning argument shows that $v_n(s,t) = \mu u_n(s,t)$. Now let us consider the first term in the bound (8.3).

$$\begin{aligned}
&\mathbb{E}\Big|\sum_{\lambda} \mathcal{I}_\lambda \,\mathbb{E}\big(L - L^\lambda \mid \mathbf{W}\big) - \sigma^2\Big| \\
&\le \sum_{s=0}^{k}\sum_{t=0}^{k} \mathbb{E}\,\Big| \sum_{\lambda\in\mathscr{S}_n(k)} \mathcal{I}_\lambda \,\mathbb{E}(S_\lambda(s,t) \mid \mathbf{W}) - \mu u_n(s,t)\Big| \\
&\le \sum_{s=0}^{k}\sum_{t=0}^{k} (|\mathscr{S}_n(k)| \cdot \mathbb{E}\,|\mathcal{I}_{\lambda_0}\,\mathbb{E}(S_{\lambda_0}(s,t) - u_n(s,t) \mid \mathbf{W})| + |u_n(s,t)| \cdot \mathbb{E}\,|L-\mu|) \\
&\le \mu \sum_{s=0}^{k}\sum_{t=0}^{k} \mathbb{E}(|\,\mathbb{E}(S_{\lambda_0}(s,t) \mid \mathbf{W}) - u_n(s,t)| \mid \mathcal{I}_{\lambda_0}) + \sigma \sum_{s=0}^{k}\sum_{t=0}^{k} |u_n(s,t)|.
\end{aligned}$$

Similarly for the second term in (8.3) we have

$$\begin{aligned}
\sqrt{\mathbb{E}((L-L^{\lambda_0})^2 \mid \mathcal{I}_{\lambda_0})} &\le \sum_{s=0}^{k}\sum_{t=0}^{k} \sqrt{\mathbb{E}(S_{\lambda_0}(s,t)^2 \mid \mathcal{I}_{\lambda_0})} \\
&\le \sum_{s=0}^{k}\sum_{t=0}^{k} (|u_n(s,t)| + \sqrt{\operatorname{Var}(S_{\lambda_0}(s,t) \mid \mathcal{I}_{\lambda_0})}).
\end{aligned}$$

The proof of the variance estimate in Theorem 3.9 shows that $u_n(s,t) \ge 0$ for $st > 0$ and $u_n(s,t) = -|\mathscr{S}_{\lambda_0}(s,t)|p_n$ for $st = 0, s+t > 0$. In particular we have

$$\sum_{s=0}^{k}\sum_{t=0}^{k} |u_n(s,t)| = \frac{1}{\mu}\sum_{s=0}^{k}\sum_{t=0}^{k} |v_n(s,t)| \le \frac{c\sigma^2}{\mu}$$

for some constant $c > 0$. Combining, the bound (8.3) reduces to

$$\begin{aligned}
d_{\mathcal{W}}(\hat{L}, \mathrm{N}(0,1)) \le &\sum_{s=0}^{k}\sum_{t=0}^{k} \frac{\mu}{\sigma^2}\,\mathbb{E}(|\,\mathbb{E}(S_{\lambda_0}(s,t) \mid \mathbf{W}) - u_n(s,t)| \mid \mathcal{I}_{\lambda_0}) + \frac{c\sigma}{\mu} \\
&+ \left(\sqrt{\frac{c^2\sigma}{\mu}} + \sum_{s=0}^{k}\sum_{t=0}^{k} \sqrt{\frac{\mu}{\sigma^3}\operatorname{Var}(S_{\lambda_0}(s,t) \mid \mathcal{I}_{\lambda_0})}\right)^2. \qquad (8.4)
\end{aligned}$$

From (8.1) it follows that $\sigma/\mu \to 0$ as $n \to \infty$. Moreover, for $st = 0$ we have $|S_{\lambda_0}(s,t)| \le 1$ a.s. Note that

$$\frac{\sigma^2}{\mu} = n^{k-2+2/(k+1)+o(1)} \text{ and } \frac{\sigma^3}{\mu} = n^{2k-3+3/(k+1)+o(1)}.$$

Thus the case $st = 0$ is negligible and we are left to prove that

$$\begin{aligned}
\Gamma_1 &:= \sum_{s=1}^{k}\sum_{t=1}^{k} \frac{\mu}{\sigma^2}\,\mathbb{E}(|\,\mathbb{E}(S_{\lambda_0}(s,t) \mid \mathbf{W}) - u_n(s,t)| \mid \mathcal{I}_{\lambda_0}) \to 0 \\
\Gamma_2 &:= \sum_{s=1}^{k}\sum_{t=1}^{k} \sqrt{\frac{\mu}{\sigma^3}\operatorname{Var}(S_{\lambda_0}(s,t) \mid \mathcal{I}_{\lambda_0})} \to 0
\end{aligned}$$

as $n \to \infty$. Clearly

$$\mathbb{E}(|\,\mathbb{E}(S_{\lambda_0}(s,t) \mid \mathbf{W}) - u_n(s,t)| \mid \mathcal{I}_{\lambda_0}) \leq \sqrt{\operatorname{Var}(S_{\lambda_0}(s,t) \mid \mathcal{I}_{\lambda_0})}. \tag{8.5}$$

Recall that,

$$S_{\lambda_0}(s,t) := \sum_{\gamma \in \mathscr{S}_{\lambda_0}(s,t)} (\mathcal{I}_\gamma - \mathcal{I}_\gamma(\mathbf{W}^{\lambda_0}))$$

where

$$\mathscr{S}_\lambda(s,t) = \{\gamma \in \mathscr{S}_n(k) \mid |\lambda \cap \gamma| = (k-s, k-t)\}.$$

We start with the term Γ_1. We consider different cases depending on the values of s,t. Note that, $\mathbb{E}(S_{\lambda_0}(s,t) \mid \mathcal{I}_{\lambda_0}) = u_n(s,t) \ll \sigma^2/\mu$ for $st < k^2$. Thus, heuristically for $st < k^2$, the contribution in Γ_1 should be $\ll 1$ as $n \to \infty$. Obviously the nontrivial case is when $s = t = k$.

Case 1. $st > 0, s + t \leq 2k - 2$: In this case we have $u_n(s,t) \geq 0$ and thus

$$\mathbb{E}(\mathcal{I}_\gamma(\mathbf{W}^{\lambda_0}) \mid \mathcal{I}_{\lambda_0}) \leq \mathbb{E}(\mathcal{I}_\gamma \mid \mathcal{I}_{\lambda_0})$$

for $\gamma \in \mathscr{S}_{\lambda_0}(s,t)$. Now we have

$$\begin{aligned}
&\mathbb{E}(|\,\mathbb{E}(S_{\lambda_0}(s,t) \mid \mathbf{W}) - u_n(s,t)| \mid \mathcal{I}_{\lambda_0}) \\
&\leq \sum_{\gamma \in \mathscr{S}_{\lambda_0}(s,t)} \mathbb{E}(\mathcal{I}_\gamma + \mathcal{I}_\gamma(\mathbf{W}^{\lambda_0}) \mid \mathcal{I}_{\lambda_0}) + |u_n(s,t)| \\
&= \sum_{\gamma \in \mathscr{S}_{\lambda_0}(s,t)} \mathbb{E}(\mathcal{I}_\gamma + \mathcal{I}_\gamma(\mathbf{W}^{\lambda_0}) \mid \mathcal{I}_{\lambda_0}) + \sum_{\gamma \in \mathscr{S}_{\lambda_0}(s,t)} \mathbb{E}(\mathcal{I}_\gamma - \mathcal{I}_\gamma(\mathbf{W}^{\lambda_0}) \mid \mathcal{I}_{\lambda_0}) \\
&= 2\binom{k}{s}\binom{k}{t}\binom{n-k}{s}\binom{n-k}{t} \mathbb{P}(\mathcal{I}_{[s+1,s+k]\times[t+1,t+k]} \mid \mathcal{I}_{[k]\times[k]}).
\end{aligned}$$

Now using the results in case 3 and 5 from the proof of Theorem 3.9 we have

$$\mathbb{P}(\mathcal{I}_{[s+1,s+k]\times[t+1,t+k]} \mid \mathcal{I}_{[k]\times[k]}) \leq n^{-k+2k(k-s)(k-t)/(2k^2-st)+o(1)} \tag{8.6}$$

and

$$2\binom{k}{s}\binom{k}{t}\binom{n-k}{s}\binom{n-k}{t} \mathbb{P}(\mathcal{I}_{[s+1,s+k]\times[t+1,t+k]} \mid \mathcal{I}_{[k]\times[k]}) \leq \varepsilon_n \sigma^2/\mu$$

where

$$\varepsilon_n := O((\log n/n^2)^{\frac{k-1}{(k+1)(k^2+2k-1)}} (\log n)^{2k-1}).$$

Thus we have

$$\frac{\mu}{\sigma^2} \mathbb{E}(|\,\mathbb{E}(S_{\lambda_0}(s,t) \mid \mathbf{W}) - u_n(s,t)| \mid \mathcal{I}_{\lambda_0}) \leq \varepsilon_n.$$

Case 2: $s + t = 2k - 1$: This corresponds to the set of matrices which have exactly one row in common with λ_0 and no columns, or vice-vera. Without loss of generality assume the former case (the later is dealt with identically) so that $s = k-1, t = k$. By (8.5) it is enough to prove that

$$\mathbb{E}(S_{\lambda_0}(k-1,k)^2 \mid \mathcal{I}_{\lambda_0}) \ll \sigma^4/\mu^2.$$

Note that

$$\mathbb{E}(S_{\lambda_0}(k-1,k) \mid \mathcal{I}_{\lambda_0}) = v_n(k-1,k)/\mu \ll \sigma^2/\mu.$$

We will write

$$\hat{\mathcal{I}}_\gamma := \mathcal{I}_\gamma(\mathbf{W}^{\lambda_0}) \text{ and } \mathbb{P}_{\lambda_0}(\cdot) = \mathbb{P}(\cdot \mid \mathcal{I}_{\lambda_0}).$$

Note that, any matrix in $\mathscr{S}_{\lambda_0}(k-1,k)$ is contained in the sub matrix $[n] \times [k+1,n]$ with exactly one row with index in $[k]$. For two matrix indices $\gamma, \gamma' \in \mathscr{S}_{\lambda_0}(k-1,k)$ define $\mathcal{N}(\gamma,\gamma') = (\ell, r, c)$ where $\ell = 1$ if γ, γ' share a row in $[k]$ and 0 otherwise; r is the number of common rows between γ, γ' in $[k+1,n]$ and c is the number of common columns between γ, γ'. Note that

$$\begin{aligned}
&|\{(\gamma,\gamma') \mid \mathcal{N}(\gamma,\gamma') = (\ell,r,c)\}| \\
&= k((k-1)\mathbb{1}\{\ell=0\} + \mathbb{1}\{\ell=1\})\binom{n-k}{k}\binom{k}{c}\binom{n-2k}{k-c} \\
&\qquad\qquad\qquad\qquad \binom{n-k}{k-1}\binom{k-1}{r}\binom{n-2k+1}{k-1-r} \\
&= O(n^{4k-2-r-c}).
\end{aligned}$$

Thus we have

$$\begin{aligned}
&\mathbb{E}(S_{\lambda_0}(k-1,k)^2 \mid \mathcal{I}_{\lambda_0}) \\
&\le O_k(1)\sum_{\ell=0}^{1}\sum_{r=0}^{k-1}\sum_{c=0}^{k} n^{4k-2-r-c}\,\mathbb{E}((\mathcal{I}_\gamma - \hat{\mathcal{I}}_\gamma)(\mathcal{I}_{\gamma_{\ell,r,c}} - \hat{\mathcal{I}}_{\gamma_{\ell,r,c}}) \mid \mathcal{I}_{\lambda_0})
\end{aligned}$$

where $\mathcal{N}(\gamma, \gamma_{\ell,r,c}) = (\ell, r, c)$. Now for $\gamma, \gamma' \in \mathscr{S}_n(k)$ we have

$$\begin{aligned}
&\mathbb{E}((\mathcal{I}_\gamma - \hat{\mathcal{I}}_\gamma)\cdot(\mathcal{I}_{\gamma'} - \hat{\mathcal{I}}_{\gamma'}) \mid \mathcal{I}_{\lambda_0}) \\
&= \mathbb{P}_{\lambda_0}(\mathcal{I}_\gamma\hat{\mathcal{I}}_\gamma^c\mathcal{I}_{\gamma'}\hat{\mathcal{I}}_{\gamma'}^c) - \mathbb{P}_{\lambda_0}(\mathcal{I}_\gamma\hat{\mathcal{I}}_\gamma^c\mathcal{I}_{\gamma'}^c\hat{\mathcal{I}}_{\gamma'}) - \mathbb{P}_{\lambda_0}(\mathcal{I}_\gamma^c\hat{\mathcal{I}}_\gamma\mathcal{I}_{\gamma'}\hat{\mathcal{I}}_{\gamma'}^c) + \mathbb{P}_{\lambda_0}(\mathcal{I}_\gamma^c\hat{\mathcal{I}}_\gamma\mathcal{I}_{\gamma'}^c\hat{\mathcal{I}}_{\gamma'}).
\end{aligned}$$

For $rc = 0, r + c > 1$ the contribution in $\mathbb{E}(S_{\lambda_0}(k-1,k)^2 \mid \mathcal{I}_{\lambda_0})$ is bounded by $O_k(1)n^{4k-2-r-c}n^{-2k} \le O_k(1)n^{2k-4}$. To see this, consider the first term in the above equation. Here we require both γ, γ' to be locally optimal, in particular column optimal and thus must possess the largest k row sums in their respective column set, each of which has probability (even conditioning on $\mathcal{I}_{\lambda_0}$) of at most than $1/\binom{n-2k}{k}$. When $r + c = 0$, one can prove that (using the method used in the proof of Theorem 3.9 for $s = t = k$)

$$\mathbb{E}((\mathcal{I}_\gamma - \hat{\mathcal{I}}_\gamma)(\mathcal{I}_{\gamma_{\ell,r,c}} - \hat{\mathcal{I}}_{\gamma_{\ell,r,c}}) \mid \mathcal{I}_{\lambda_0}) = O(n^{-2k-2})$$

and for $r + c = 1$

$$\mathbb{E}((\mathcal{I}_\gamma - \hat{\mathcal{I}}_\gamma)(\mathcal{I}_{\gamma_{\ell,r,c}} - \hat{\mathcal{I}}_{\gamma_{\ell,r,c}}) \mid \mathcal{I}_{\lambda_0}) = O(n^{-2k-1}).$$

The n^{-2k} term comes from the probability that both γ and $\gamma_{\ell,r,c}$ are locally optimal and the $1/n$ improvement is coming from the fact that $\mathbb{E}(\mathcal{I}_\gamma - \hat{\mathcal{I}}_\gamma \mid \mathcal{I}_{\lambda_0}) = O(n^{-k-1})$. Thus for $rc = 0$, the total contribution is $O(n^{2k-4})$. When $rc \in [1, k(k-1)], \ell \in \{0,1\}$, the contribution is

$$n^{4k-2-r-c}n^{-2k+\frac{2k(r+\ell)c}{2k^2-(k-r-\ell)(k-c)}}.$$

The maximum power occurs for $r = c = \ell = 1$ so that the contribution is bounded by

$$n^{2k-4+\frac{2k(1+\ell)}{2k^2-(k-1-\ell)(k-1)}} \le n^{2k-4+4k/(k^2+3k-2)}$$

and $4k/(k^2+3k-2) = 4/(k+1) - 8(k-1)/((k+1)(k^2+3k-2))$. Thus combining everything we have

$$\begin{aligned}\mathbb{E}(S_{\lambda_0}(k-1,k)^2 \mid \mathcal{I}_{\lambda_0}) &= O(n^{2k-4+4/(k+1)-8(k-1)/((k+1)(k^2+3k-2))})\\ &= O(n^{-8(k-1)/((k+1)(k^2+3k-2))}\cdot \sigma^4/\mu^2).\end{aligned}$$

Case 3. $s=t=k$. This corresponds to the set of matrices which have no common rows or columns with λ_0. We move to the proof of

$$\operatorname{Var}(S_{\lambda_0}(k,k) \mid \mathcal{I}_{\lambda_0}) \ll \sigma^4/\mu^2.$$

Note that, any matrix in $\mathscr{S}_{\lambda_0}(k,k)$ is contained in the sub matrix $[k+1,n]\times[k+1,n]$. Also we have

$$\begin{aligned}&|\{\gamma,\gamma'\in\mathscr{S}_{\lambda_0}(k,k) \mid |\gamma\cap\gamma'|=(r,c)\}|\\ &=\binom{n-k}{k}\binom{k}{r}\binom{n-2k}{k-r}\binom{n-k}{k}\binom{k}{c}\binom{n-2k}{k-c} = O(n^{4k-r-c}).\end{aligned}$$

Thus we have

$$\begin{aligned}&\operatorname{Var}(S_{\lambda_0}(k,k)^2 \mid \mathcal{I}_{\lambda_0})\\ &\le O(1)\sum_{\ell=0}^{1}\sum_{r=0}^{k-1}\sum_{c=0}^{k} n^{4k-r-c}\operatorname{Cov}(\mathcal{I}_\gamma-\hat{\mathcal{I}}_\gamma, \mathcal{I}_{\gamma_{r,c}}-\hat{\mathcal{I}}_{\gamma_{r,c}} \mid \mathcal{I}_{\lambda_0})\end{aligned}$$

where $\mathcal{N}(\gamma,\gamma_{r,c})=(r,c)$. For $r=c=k$, the total contribution in the variance is

$$O(n^{2k}n^{-k-1}) \le O(n^{2k-4+3/(k+1)}).$$

Note that here n^{-k-1} term comes from the fact that $\mathcal{I}_\gamma$ has probability n^{-k} and after changing the elements in the first k rows and k columns γ is no longer locally optimal implies one of the new rows or columns beat γ which has probability $1/n$. In particular, similar to the variance calculation for L_n, for all $rc=0, r+c>1$ the contribution is

$$\le n^{4k-r-c-2k-2} \le n^{2k-4}.$$

and for all $rc\ge 1$ the contribution is

$$\begin{aligned}&n^{4k-r-c}n^{-2k+2krc/(2k^2-(k-r)(k-c))-2/(1+\max\{r,c\}/k)}\\ &\le n^{-2(k-1)/((k+1)(k^2+2k-1))}\sigma^4/\mu^2\end{aligned}$$

where the largest exponent occurs for $r=c=1$. Thus the only terms remaining to bound are when $r+c=1$ and $r+c=0$. We look at the $r+c=0$ case first. We want to bound

$$\sum_{\gamma,\gamma'\in\mathscr{S}_{\lambda_0}(k,k),|\gamma\cap\gamma'|=(0,0)}\operatorname{Cov}(\mathcal{I}_\gamma-\hat{\mathcal{I}}_\gamma,\mathcal{I}_{\gamma'}-\hat{\mathcal{I}}_{\gamma'} \mid \mathcal{I}_{\lambda_0}).$$

Number of summands in the above sum is $O(n^{4k})$. Now after some simplification it is easy to see that we need to bound

$$\operatorname{Cov}(\mathcal{I}_\gamma\hat{\mathcal{I}}^c_\gamma,\mathcal{I}_{\gamma'}\hat{\mathcal{I}}^c_{\gamma'} \mid \mathcal{I}_{\lambda_0})$$

which, by Lemma 3.10 can be bounded by

$$n^{-2k-2-2k/(k+1)} = n^{-2k-4+2/(k+1)}.$$

Thus the total contribution is

$$n^{4k-2k-4+2/(k+1)} = n^{2k-4+4/(k+1)-2/(k+1)} = n^{-2/(k+1)}\sigma^4/\mu^2.$$

Similarly for the $r = 1, c = 0$ case the total contribution is

$$n^{4k-1}n^{-2k-2-1} = n^{2k-4} = n^{-4/(k+1)}\sigma^4/\mu^2.$$

Combining everything we have $\Gamma_1 \to 0$ as $n \to \infty$.

Now we show that

$$\Gamma_2 = \sum_{s=1}^{k}\sum_{t=1}^{k}\sqrt{\frac{\mu}{\sigma^3}\operatorname{Var}(S_{\lambda_0}(s,t) \mid \mathcal{I}_{\lambda_0})} \to 0$$

as $n \to \infty$. Note that, $\mathbb{E}(S_{\lambda_0}(s,t) \mid \mathcal{I}_{\lambda_0}) = u_n(s,t) \leq \sigma^2/\mu$ for all s, t. Heuristically for fixed s, t the contribution in Γ_2 should be $\leq \sqrt{\mu/\sigma^3 \cdot \sigma^4/\mu^2} = \sqrt{\sigma/\mu} \to 0$ as $n \to \infty$. We leave the proof to the interested reader where the proof follows exactly the same steps used in case $1-3$ of the proof of $\Gamma_1 \to 0$. Combining everything finally we have the result that

$$d_{\mathcal{W}}(\hat{L}, \mathrm{N}(0,1)) \to 0 \tag{8.7}$$

as $n \to \infty$. ■

Acknowledgments. PD is grateful for the hospitality of the Department of Statistics and Operations research, University of North Carolina, Chapel Hill, where much of the research was done. SB was partially supported by NSF grant DMS-1105581. AN was partially supported by NSF grant DMS-0907177.

References

[1] L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi, *On combinatorial testing problems*, Ann. Statist. **38** (2010), no. 5, 3063–3092. MR2722464 (2011k:62035)

[2] E. Aidekon, *Convergence in law of the minimum of a branching random walk*, arXiv preprint arXiv:1101.1810 (2011).

[3] D. J. Aldous, C. Bordenave, and M. Lelarge, *Dynamic programming optimization over random data: The scaling exponent for near-optimal solutions*, SIAM Journal on Computing **38** (2009), no. 6, 2382–2410.

[4] N. Alon, M. Krivelevich, and B. Sudakov, *Finding a large hidden clique in a random graph*, Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 1998), 1998, pp. 594–598. MR1642973 (99e:68114)

[5] E. Arias-Castro, E. J. Candès, and A. Durand, *Detection of an anomalous cluster in a network*, Ann. Statist. **39** (2011), no. 1, 278–304. MR2797847 (2012a:62130)

[6] E. Arias-Castro, E. J. Candès, H. Helgason, and O. Zeitouni, *Searching for a trail of evidence in a maze*, Ann. Statist. **36** (2008), no. 4, 1726–1757. MR2435454 (2010h:62025)

[7] S. M. Berman, *Limit theorems for the maximum term in stationary sequences*, Ann. Math. Statist. **35** (1964), 502–516. MR0161365 (28 #4572)

[8] B. Bollobás and P. Erdős, *Cliques in random graphs*, Math. Proc. Cambridge Philos. Soc. **80** (1976), no. 3, 419–427. MR0498256 (58 #16408)

[9] B. Bollobás, *Random graphs*, Second, Cambridge Studies in Advanced Mathematics, vol. 73, Cambridge University Press, Cambridge, 2001. MR1864966 (2002j:05132)

[10] C. Butucea and Y. I. Ingster, *Detection of a sparse submatrix of a high-dimensional noisy matrix*, arXiv preprint arXiv:1109.0898 (2011).

[11] L. H. Y. Chen, L. Goldstein, and Q.-M. Shao, *Normal approximation by Stein's method*, Probability and its Applications (New York), Springer, Heidelberg, 2011. MR2732624 (2012b:60103)

[12] L. H. Y. Chen and Q.-M. Shao, *Stein's method for normal approximation*, An introduction to Stein's method, 2005, pp. 1–59. MR2235448

[13] Y. Dekel, O. Gurel-Gurevich, and Y. Peres, *Finding hidden cliques in linear time with high probability*, arXiv preprint arXiv:1010.2997 (2010).

[14] P. Diaconis and S. Holmes (eds.), *Stein's method: expository lectures and applications*, Institute of Mathematical Statistics Lecture Notes—Monograph Series, 46, Institute of Mathematical Statistics, Beachwood, OH, 2004. Papers from the Workshop on Stein's Method held at Stanford University, Stanford, CA, 1998. MR2118599 (2005i:62008)

[15] R. Durrett and V. Limic, *Rigorous results for the NK model*, Ann. Probab. **31** (2003), no. 4, 1713–1753. MR2016598 (2005a:60067)

[16] S. N. Evans and D. Steinsaltz, *Estimating some features of NK fitness landscapes*, Ann. Appl. Probab. **12** (2002), no. 4, 1299–1321. MR1936594 (2004b:60131)

[17] S. Fortunato, *Community detection in graphs*, Physics Reports **486** (2010), no. 3, 75–174.

[18] J. Galambos, *On the distribution of the maximum of random variables*, Ann. Math. Statist. **43** (1972), 516–521. MR0298730 (45 #7779)

[19] M. Jerrum, *Large cliques elude the Metropolis process*, Random Structures Algorithms **3** (1992), no. 4, 347–359. MR1179827 (94b:05171)

[20] S. A. Kauffman and E. D. Weinberger, *The nk model of rugged fitness landscapes and its application to maturation of the immune response*, Journal of theoretical biology **141** (1989), no. 2, 211–245.

[21] M. R. Leadbetter, G. Lindgren, and H. Rootzén, *Extremes and related properties of random sequences and processes*, Springer Series in Statistics, Springer-Verlag, New York, 1983. MR691492 (84h:60050)

[22] W. V. Li and Q.-M. Shao, *A normal comparison inequality and its applications*, Probab. Theory Related Fields **122** (2002), no. 4, 494–508. MR1902188 (2003b:60034)

[23] V. Limic and R. Pemantle, *More rigorous results on the Kauffman-Levin model of evolution*, Ann. Probab. **32** (2004), no. 3A, 2149–2178. MR2073188 (2005b:92032)

[24] S. C. Madeira and A. L. Oliveira, *Biclustering algorithms for biological data analysis: a survey*, Computational Biology and Bioinformatics, IEEE/ACM Transactions on **1** (march 2004jan.), no. 1, 24 –45.

[25] M. W. Mahoney, *Algorithmic and statistical perspectives on large-scale data analysis*, arXiv preprint arXiv:1010.1609 (2010).

[26] M. Mézard and A. Montanari, *Information, physics, and computation*, Oxford Graduate Texts, Oxford University Press, Oxford, 2009. MR2518205 (2010k:94019)

[27] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin glass theory and beyond*, World Scientific Lecture Notes in Physics, vol. 9, World Scientific Publishing Co. Inc., Teaneck, NJ, 1987. MR1026102 (91k:82066)

[28] B. Pittel′, *On the probable behaviour of some algorithms for finding the stability number of a graph*, Math. Proc. Cambridge Philos. Soc. **92** (1982), no. 3, 511–526. MR677474 (83k:68064)

[29] C. M. Reidys and P. F. Stadler, *Combinatorial landscapes*, SIAM review **44** (2002), no. 1, 3–54.

[30] N. Ross, *Fundamentals of Stein's method*, Probab. Surv. **8** (2011), 210–293. MR2861132 (2012k:60079)

[31] A. A. Shabalin, V. J. Weigman, C. M. Perou, and A. B. Nobel, *Finding large average submatrices in high dimensional data*, The Annals of Applied Statistics **3** (2009), no. 3, 985–1012.

[32] J. M. Steele, *Probability theory and combinatorial optimization*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 69, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997. MR1422018 (99d:60002)

[33] C. Stein, *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables*, Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory, 1972, pp. 583–602. MR0402873

[34] X. Sun and A. B. Nobel, *On the maximal size of large-average and anova-fit submatrices in a gaussian random matrix*, Arxiv preprint arXiv:1009.0562 (2010).

[35] E. D. Weinberger, *Local properties of kauffman's nk model: A tunably rugged energy landscape*, Physical Review A **44** (1991), no. 10, 6399.

[36] R. Willink, *Bounds on the bivariate normal distribution function*, Comm. Statist. Theory Methods **33** (2004), no. 10, 2281–2297. MR2104113

[37] S. Wright, *The roles of mutation, inbreeding, crossbreeding and selection in evolution*, Proceedings of the sixth international congress on genetics, 1932, pp. 356–366.